PROCEEDINGS

International Conference on

Big Data, Knowledge and Control Systems Engineering - BdKCSE'2014

5th November 2014
108Rakovski Str., Hall 2, 1000 Sofia, Bulgaria

Institute of Information and Communication Technologies
- Bulgarian Academy of Sciences
John Atanasoff Society of Automatics and Informatics

Editor:

Rumen D. Andreev

Department of Communication Systems and Services
Institute of Information and Communication Technologies - Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 2, 1113 Sofia, Bulgaria

Table of contents

Organized by:

Institute of Information and Communication Technologies
- Bulgarian Academy of Sciences

John Atanasoff Society of Automatics and Informatics

Program committee

Honorary Chairs

- Acad. Vassil Sgurev — Bulgarian Academy of Sciences — Bulgaria
- Prof. John Wang — Montclair State University — USA
- Corr. Memb. Mincho Hadjiski — Bulgarian Academy of Sciences — Bulgaria

Conference Chairs

- Chairman – Rumen Andreev — Bulgarian Academy of Sciences — Bulgaria
- Vice chairman – Lyubka Doukovska — Bulgarian Academy of Sciences — Bulgaria
- Vice chairman – Yuri Pavlov — Bulgarian Academy of Sciences — Bulgaria

Program Committee

- Abdel-Badeeh Salem — Ain Sham University — Egypt
- Chen Song Xi — Iowa State University — USA
- Dimiter Velev — University of National and World Economy — Bulgaria
- Evdokia Sotirova — University "Prof. Asen Zlatarov" — Bulgaria
- František Čapkovič — Slovak Academy of Sciences — Slovakia
- George Boustras — European University — Cyprus
- Georgi Mengov — University of Sofia — Bulgaria
- Ivan Mustakerov — IICT, Bulgarian Academy of Sciences — Bulgaria
- Ivan Popchev — IICT, Bulgarian Academy of Sciences — Bulgaria
- Jacques Richalet — France
- Kosta Boshnakov — University of Chemical Technology and Metallurgy, Bulgaria
- Krasen Stanchev — Sofia University — Bulgaria
- Krasimira Stoilova — IICT, Bulgarian Academy of Sciences — Bulgaria
- Ljubomir Jacić — Technical College Požarevac — Serbia
- Ljupco Kocarev — Macedonian Academy of Sciences and Arts — Macedonia
- Milan Zorman — University of Maribor — Slovenia
- Neeli R. Prasad — Aalborg University, Princeton — USA
- Olexandr Kuzemin — Kharkov National University of Radio Electronics, Ukraine
- Peđa Milosavljević — University of Niš — Serbia
- Peter Kokol — University of Maribor — Slovenia
- Radoslav Pavlov — IMI, Bulgarian Academy of Sciences — Bulgaria
- Rumen Nikolov — UniBIT-Sofia — Bulgaria
- Silvia Popova — ISER, Bulgarian Academy of Sciences — Bulgaria
- Song II-Yeol — Drexel University — USA
- Sotir Sotirov — University "Prof. Asen Zlatarov" — Bulgaria
- Svetla Vassileva — ISER, Bulgarian Academy of Sciences — Bulgaria
- Tomoko Saiki — Tokyo Institute of Technology — Japan
- Uğur Avdan — Anadolu University — Turkey
- Valentina Terzieva — IICT, Bulgarian Academy of Sciences — Bulgaria
- Valeriy Perminov — National Research Tomsk Polytechnic University, Russia
- Vassia Atanassova — IICT, Bulgarian Academy of Sciences — Bulgaria
- Vera Angelova — IICT, Bulgarian Academy of Sciences — Bulgaria
- Vyacheslav Lyashenko — Kharkov National University of Radio Electronics, Ukraine
- Wojciech Piotrowicz — University of Oxford — UK
- Zlatogor Minchev — IICT, Bulgarian Academy of Sciences — Bulgaria
- Zlatolilia Ilcheva — IICT, Bulgarian Academy of Sciences — Bulgaria

Problems of the Big Data and Some Applications

Vassil Sgurev, Stanislav Drangajov
Institute of Information and Communication Technologies – BAS
G. Bonchev, str, 113 Sofia, Bulgaria
e-mail: vsgurev@gmail.com; sdrangajov@gmail.com

Abstract: We are trying in this paper to answer the question: What is Big DataB and how they could be used. It turns out by the investigations that this is a very wide and perspective area of research and even a special United Nations' program exists for their usage for prediction of eventual events with a great degree of probability for happening, and as so – preventive actions to be undertaken. Big Data is usually considered as an aggregate of a great bulk of data, usually unstructured and the methods of their processing for the purpose of extracting from them useful information. Several examples are shown for their usage in business and the public sector which demonstrate the profit of their usage
Keywords: Big Data, Predictive Analytics, Unstructured data

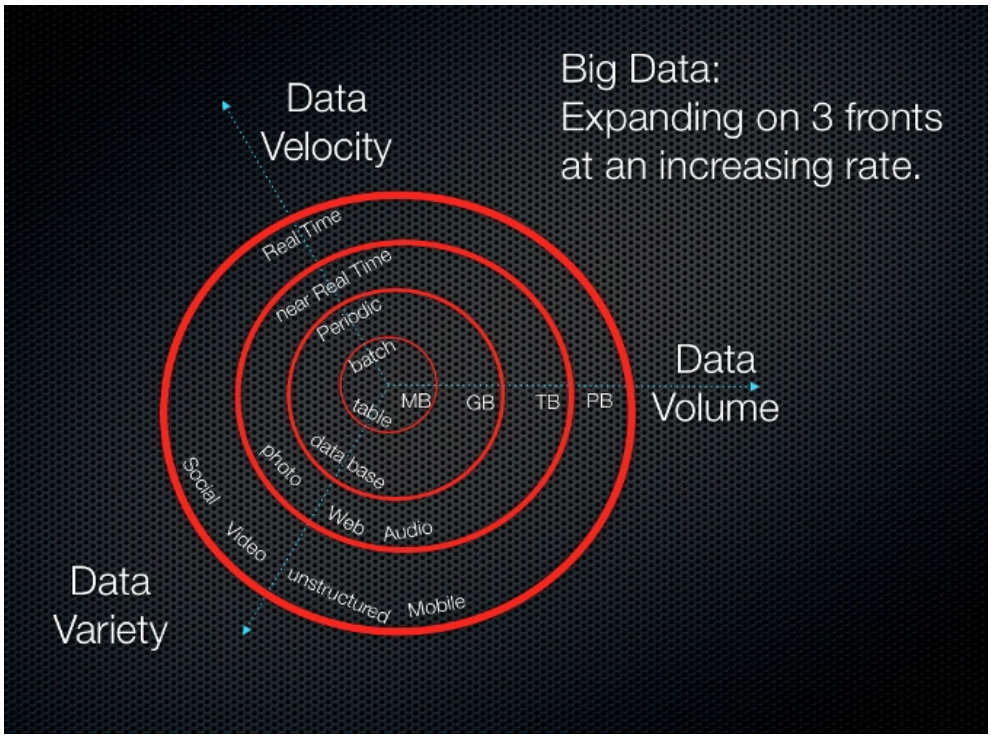
1. What is Big Data?

Big Data is a term that evolved some 10-15 years ago in the sphere of the information Technologies (IT) which seems to grow overwhelming in the modern world. The term is a buzzword and many, even we, don't want to admit that we have no exact and generally acknowledged definition. From IBM they say: "Don't focus on what Big Data is, focus on what problems Big Data can address. IBM can help you gain value from Big Data.". It is clear that Big Data are digital, but of what type? When saying big databases it is clear and they can be classified. Besides, the information there is strictly structured. But what Big Data is – volume or technology [1]? Let's assume we consider the volume – peta, exa, and so on, bytes[1]. Even today, if you wrote in any of the big web search engines e.g. Google, Yahoo!, Yandex etc. whatever comes to your mind, say for example "… in a cavern in a canyon[2].." information will be immediately delivered. This supposes more than peta, exa, zetta etc. bytes to be processed and information of thousands items to be returned in the frame of a half

[1] Peta -10^{15}, exa -10^{18}, zetta - 10^{21}, yotta - 10^{24}.
[2] Two words from a popular song of the American gold rush.

a second. Naturally this is not the case. Evidently the case in point is Big Data. We have to do nothing else except to accept that: “Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.“ [2].

It is accepted that Big Data may be characterized by the three Vs: huge Volume, tremendous Variety of data of various types to be processed, and the Velocity at which these data must be processed. All three V-s expand with greater and greater speed. It is very hard, and practically – impossible, data to be integrated by the traditional methods. For this purpose new approaches are sought, like artificial intelligence, neural networks, fuzzy logic, different types of statistical and probabilistic analysis etc. and they are really successful in many cases of data and knowledge extraction.



2. More and more jumbled

Data grow more and more jumbled. And it is really so. Some features are described in [3] which may be useful for the explanation of Big Data and their possible usage. The first one concerns the volume of the data accessible. Researchers and analysts search traditionally data immediately related to their investigations. Modern technologies provide gathering and

analysis of a huge amount of data, although indirectly related to a given investigation. Naturally by increasing the scale of the information inaccuracies increase also.

The second characteristic is that by enlarging the storage and accessibility of data of more and more data the latter grow messier as there are no criteria for their validity. But the volume of data supplies more valuable information, despite of the more errors that may exist.

A definition of Big Data by IDC [4], although too broad, looks like acceptable tieing volume and technologies. “Big Data technologies describe a new generation of technologies and architectures, designed so organizations like yours can economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. This world of Big Data requires a shift in computing architecture so that customers can handle both the data storage requirements and the heavy server processing required to analyze large volumes of data economically.”

3. Correlations vs. causality.

The third characteristic is the trend to passing from causality to correlation connection. It is generally accepted in scientific research that correlation does not prove a causality connection. New data mining techniques may provide information about WHAT is happening, and not –WHY it is happening. Correlations give relationship between data. Quite sophisticated methods are used in this case – statistical, probabilistic, artificial intelligence, fuzzy logic, neural networks etc. This, no matter how strange it seems, provides an opportunity for predictive analytics which is one of the strong aspects of Big Data. They may be used in many spheres of business management and social relations. We will point out some amazing examples from the real world in the next item.

4. Predictive analysis, examples

A furious man rushed into a supermarket of a big US supermarket chain very angry that they send his daughter, a teenager, advertisement messages related to pregnancy and babies. Why does the chain encourage pregnancy of teenager girls!? The supermarket manager was not aware of the case and what had happened and apologized to the angry parent. Several days later the manager decided to check and follow up the case by a phone call. The father answered very embarrassed. It turned out his daughter was really pregnant and the father admitted that he had not been aware of the whole truth.

It turned out that in the supermarket chain they recently have started to use Big Data techniques to analyze the huge volume of data collected to send more personalized

advertisements to customers. Tracking their shopping habits analysts encountered interesting correlations:

„Women on the baby department were buying larger quantities of un scented lotion around the beginning of their second trimester. Another analyst noted that that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. And so on …"

Knowing this data and acting with respect to it the supermarket chain sends customized advertisements to women who according to their analysis are pregnant. This is the way they knew that the girl is pregnant before her patents. This narrative demonstrates how unstructured data may be used in the complex analysis of the customers' demand.

Another very interesting and of great social importance area is the social security. This is also known as predictive policing. This is used for criminological models of data and data from past criminal and even terroristic cases to predict the probabilities where and when new crimes may occur. This is useful for the operational and strategic actions of police, e.g. patrol routes, showing of possible "hot points" etc. An example: During the Boston marathon bombing in 2013 police adopted a new approach to data gathering in their investigation. They used the so called "crowd sourcing". This is a technique of gathering information about new products and services by commercial sites. Boston police appealed anybody who had picture or video of the event to send them a copy. It is clear this is much more efficient than phone calls, descriptions etc. And in the end it turned out that through these thoroughly unstructured and heterogeneous data police identified end caught the assailants in very short time.

Police gathers daily huge amounts of data, both for operational and investigative purposes. With the time these data may produce a picture of the development of the criminal and of the police actions in this direction and how police manages with its tasks. Over time these data increase and it gets more important they to be used in decision making.

After the terror attacks in Norway in 2011 authorities there, police and intelligence service, were ruthlessly criticized, that they had not used to a sufficient degree the data analysis. The Norwegian Board of Technology (NBT) immediately started a project named "Openness and Security after the 22nd of July". The project manager explains the idea behind the project "By feeding criminological models with both crime data and data from other sources, computers can calculate where and when are likely to happen. The predictions can be surprisingly precise, allowing the police to be on the site before anything actually happens."

In Bulgaria analogical statistical methods are of course used for more than a century, the Statistical Bureau being one of the first institutions of the modern Bulgarian state, but as far as we are informed the Big Data techniques are not widely used for predictive analytics. This approach seems to be very useful for predicting and avoiding heavy industrial and transport accidents, even natural calamities and terroristic actions.

5. Privacy in the time of Big Data

The issue of privacy is really very ticklish in the virtual computer world and may be it is impossible to find a definitive answer. When we gain something like free access to the virtual space, no doubt we must pay the price. It is practically impossible someone to remain absolutely anonymous in the information space even through a smart phone call. In the case of Big Data the situation is analogical. From Microsoft they propose „Differential Privacy". In this approach from Microsoft propose a "small" inaccuracies of the information containing personal data but allows downloading information from big databases. All this is OK but anyway there is SOMEONE who has the whole authentic information, IP address, location etc. You see how the most sacred documents of the biggest secret and intelligence services become generally known. May be this is the sacrifice that society must make to use the profit of big data. Here we do not discuss encapsulated societies where free access to information is prohibited. That is why everyone must personally decide what, how much, and whom to provide with personal information. Some legislative regulation of the issue of personal data manipulation is unconditionally needed, but this is not the subject of this paper.

6. Global Pulse

This is an initiative of the UN General Secretary launched in 2009. The initiative is described in details in its portal [5]. It is focused on using Big Data techniques for monitoring and tracking the impacts of local, or even global scale, socio-economic crises and predicting with good reliability their possible arising. We quote below the official definition of the initiative according to the UN site.

„Global Pulse is a flagship innovation initiative of the United Nations Secretary-General on big data. Its vision is a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action.

The initiative was established based on a recognition that digital data offers the opportunity to gain a better understanding of changes in human well-being, and to get real-time feedback on how well policy responses are working."

Some of the main partners of the initiative are Telenor, a very big international telecommunication company, and the Harvard University. One of the senior officers of Telenor says that they collect huge volumes of data and as a commercial company Telenor uses this data for product development and advertisement. But we would be inspired to provide some of these data for humanitarian projects.

As an example an investigation is pointed out how the movement of people affects the outbreak and spread of diseases in Asian countries. Big Data techniques are used and as a result some general recommendations are given to the government in which regions preventive measures should be previously undertaken to avoid the disaster, epidemics, etc.

A similar project was carried out in Kenya for the purpose of limiting the malaria and due to it researchers recommended the government to take precautions around the Victoria Lake as one of the most active "hubs" in the malaria dissemination. As a result the by eliminating the disease in this area would lead to much fewer outbreaks in other areas.

This is an evident demonstration that Big Data represent a global world interest and find practical application, and they are not a research exercise only.

7. Summing up

An attempt is made in the work the concept of big data to be clarified on the base of information from sources which are leading in this area. It is pointed out that under Big Data one can understand the aggregate of huge in volume date, their velocity of propagation and their variety, together with the methods with their processing and retrieval of the useful information. Some examples are given for their practical application in business, health care, social life etc. Each organization may perform analysis of its activity and take advantage of the possibilities that the Big Data concept offers.

References

[1] Vangie Beal, http://www.webopedia.com/TERM/B/big_data.html

[2] Margaret Rouse, http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

[3] Viktor Mayer-Schönberger and Kenneth Cukier, BIG DATA:A Revolution That Will Transform How We Live, Work, and Think (2013), Dolan/Houghton Mifflin Harcourt (HMH)

[4] Richard L. Villars Carl W. OlofsonMatthew Eastwood
http://sites.amd.com/us/Documents/IDC_AMD_Big_Data_Whitepaper.pdf
[5] http://www.unglobalpulse.org/
[6] http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[7] http://www.bb-team.org/articles/4921_korelaciata-ne-dokazva-prichinno-sledstvena-vruzka

Estimation of Flood Risk Zones of Maritza River and its Feeders on the Territory of Svilengrad Municipality as Part of Smart Water Project WEB-GIS Tool

Nina Dobrinkova, Valentin Slavov
Institute of Information and Communication Technologies – Bulgarian Academy of Sciences,
acad. Georgi Bonchev str. bl. 2, 1113 Sofia, Bulgaria
e-mail: nido@math.bas.bg, val_slavov@mail.bg

Abstract: The presented paper will focus on flood risk mapping on the territory of Svilengrad municipality, where Maritza River and its feeders are causing huge flood events during the spring season. The high wave evaluation and its implementation in the web-GIS tool, part of the Smart Water project supported under DG "ECHO" call for prevention and preparedness, will give illustration of the first attempts of application of INSPIRE directive on the Bulgarin-Turkish-Greek border zone.
Keywords: Smart Water project, Flood Risk Mapping, Svilengrad Municipality, hydrological estimation of "high" waves.

1. Introduction

Floods can create damages and human casualties with high negative impact for the society. Some of the most devastating floods have happened in Europe in the last ten years. In response EU has accepted a Directive 2007/60/EC, which the European Parliament and Council of the European Union published on October 23, 2007 with scope about the assessment and management of flood risks [1]. The Directive establishes a framework for assessment and management of flood risks to reduce the associated effects on human health, environment, cultural heritage and economic activities. With accordance to the Directive the Bulgarian Executive Agency of Civil Defense (EACD) has introduced categories of floods, depending on their size, frequency and duration [2].

In this paper will be shown structured hydrologic estimation of the high flows, formed after intensive precipitations in Maritza River and its feeders on the territory of Svilengrad and neighboring municipalities. The computed results will be implemented in the web-GIS tool, which structure will be also presented as modules for Civil Protection Response capacity support for decision making by the responsible authorities.

1.1 Feeders of Maritza River in the region of Svilengrad Municipality

Maritza is the biggest Bulgarian river. The river is subject of observations in Svilengrad since 1914, during the period 1914-1972 they have been made on the stone bridge of the river built in 15th century (cultural heritage of UNESCO), and after 1972 the observations are realized on the new railway bridge. Since 1990 the observations have been accomplished on the new highway bridge.

The references provide the following data about the orohydrographic characteristics of the river catchment up to the hydrometric point.

The subject considered is all feeders of Maritza River on the territory of Svilengrad municipality. These feeders influence directly the formation of high flows and their accounting guarantees safe exploitation of the existing protective installations close to the river. The most important feeders of Maritza River with a catchment area above 8 sq. km in the order of their influx are:

- Siva river – a right feeder on the boundary with Ljubimets municipality;
- Mezeshka river – a right feeder;
- Goljiamata reka (Kanaklijka) – a left feeder;
- Levka river – a left feeder;
- Selska reka – a left feeder;
- Jurt dere – a left feeder;
- Tolumba dere – a left feeder;
- Kalamitza – a left feeder.

The data about the catchment areas of the feeders and of Maritza river itself, given in the next table, are defined from topographic maps in scale 1:50 000 and 1:25 000.

No	Name	Feeder	Area Sq. km	Altitude m
1	2	3	4	5
1	Siva river	right	28,512	359
2	Mezeshka river	right	34,716	315
3	Maritza		20840,00	582,00
4	Goljiamata rekar	left	171,762	399
5	Levka	left	144,121	449
6	Selskata reka	left	38,424	230
7	Kalamitza	left	65,437	211,25

Table 1: Maritza River feeders bigger than 8 sq. km. catchment area

Among the rivers with a catchment area above 8 sq. km, there are smaller rivers and ravines, as well as areas that directly outflow to Maritza River. The small rivers are shown in the table given below.

As a whole, for all small catchments above mentioned, it could be accepted that they have altitude of about 60-90 m, average terrain slope 0,10-0,11 and forestation of 15-20 %.

No	Name	On the land of	Area sq. km
1	2	3	4
1	Total for village Momkovo.	Village Momkovo	**4,20**
2	Total for district"Novo selo"	District"Novo selo"	**3,18**
3	Total for Svilengrad	Svilengrad town	**4,83**
4	Total for village Generalovo	village Generalovo	**2,06**
5	Ravine "Jurt dere"	Village Captain Andreevo	**4,94**
6	Ravine "Tolumba dere"	Village Captain Andreevo	**3,32**
7	Total for village Captain Andreevo	Village Captain Andreevo	**3,10**

Table 2: Catchments of small rivers and ravines that outflow to Maritza River on the territory of Svilengrad municipality

For calculation purposes the climatic characteristics of the municipalities from where the Maritza River feeders flow are also presented.

1.2 Climatic characteristics

For the present research, important are only the rainfalls, form the runoff and the high flow. The distribution of the rainfalls during the year determines the transitional climatic character of the Thracian lowland, namely, with summer and winter rainfall peaks. In Tables 3, 4 and 5 data is given, prepared according to the records from the National Institute of Metrology and Hydrology with the main characteristics of the rainfalls provided for seven hydro-metric stations (HMS) at Svilengrad, Topolovgrad, Elhovo, Haskovo, Harmanli, Opan, Lyubimetz.

HMS	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	Aver.
Svilengrad	63	46	38	49	57	58	36	26	34	56	63	69	696
Topolovgrad	62	49	42	52	58	63	47	32	37	55	69	72	637
Elhovo	46	42	35	45	53	58	43	28	36	44	59	56	547
Haskovo	63	47	50	57	67	69	40	37	34	61	67	75	668
Harmanli	52	37	36	49	56	62	39	29	35	53	64	65	576
Opan	50	39	38	47	62	64	46	35	28	50	55	58	571
Lyubimetz	53	40	36	47	58	57	36	23	33	53	60	63	559

Table 3: Average long-term rainfall monthly amounts in mm

Station	24-hour monitoring (maximum)		Probability in %							
	mm	year	2%	5%	10%	25%	50%	75%	90%	95%
Haskovo	103,7	1932	98	85	75	60	46	36	30	28
Harmanli	115	1984	90	78	68	53	39	29	23	20
Lyubimetz	88	1984	78	69	61	49	38	30	25	22
Svilengrad	97,4	1969	84	74	66	54	42	34	28	26
Topolovgrad	123,1	1940	107	91	79	61	44	33	26	24
Galabovo	122,0	1962	97	82	70	53	39	29	23	22

Table 4: The maximum diurnal rainfalls through the years with different probability

During the period 1976-1983, the following parameters of the maximum rainfalls and their values for different security rates were used for the representative stations as given below:

Station	Altitude	24-hour monitoring.	$N_{макс,абс}$	$N_{макс,ср}$			Probability in %			
					Cv	Cs	0,1%	0,5%	1%	2%
	м	In years	мм	мм						
Izvorovo village	350	46		48	0,48	1,92	191	144	127	105
Ravna gora village	380	23		48	0,40	1,60	150	117	106	90
Haskovo bani	390	25	91	49	0,38	1,52	156	120	111	102
Haskovo	180	81	104	52	0,35	1,4	154	122	111	100
Elena village	210	37	108	52	0,40	1,6	176		120	
Cv. Poliana	210		96	50	0,27	1,08	115		91	
Byagovo	235	48	87	49	0,33	1,32	137	111	101	94
Galabovo	104	41	106	46,75	0,45	1,8	172	133	118	107

Table 5: The maximum diurnal rainfalls in the different measurement stations during the years

From the analysis of the data was established that at the hydrological assessment of the micro-dams data with higher than the maximum values of the rainfalls with different security rate were used. For security purposes it can be assumed that these values will better guarantee the trouble-free functioning of the facilities, and for this reason, further on this data was used for the maximal rainfalls. By using the received values of the high flow with different security rates the values of the maximum runoff with different security were

calculated, through which the regional dependence was established and used for determination of the runoff as formed from the smaller additional catchment areas.

1.3 Modulus and norm of the runoff

The modulus of the river runoff was determined through the creation of a regional dependence between the runoff modulus and the average sea level of the water catchment basins of the rivers within the region.

№	Name	Area (km^2)	Average above sea level (m)	Outflow module (l/s/sq.km)	Outflow rate (l/s) $Q=M.F$
1	2	3	4	5	6
1	Siva river	28,512	359,00	5,5691	158,788
2	Mezeshka river	34,710	315,00	5,3060	184,172
3	Golyama rekar	171,762	399,00	5,8197	999,597
4	Levka river	144,121	449,00	6,1487	886,158
5	Selska reka	38,424	230,00	4,8324	185,680
6	Kalamitza river	65,437	211,25	4,7337	309,762

Table 6: Calculated liters which are potential threat along the rivers listed in the table

According to the regulations, the culverts and bridges of the railroad should be designed for security rate of 1%, thus it is assumed that the surrounding territory is threatened once per one hundred years. It is checked whether the correction with the security rate foreseen is able to accept the high flow with a security rate of 0,1% - i.e., one thousand years/flow (wave).

Due to the lack of direct measurements, the maximum quantities with different security rates were determined by indirect approximate methods. A comparatively reliable value of the maximum water quantity can be received by the so-called river-bed method, where by Shezi's hydraulic formula with a maximum river water level, the maximum flow rate is calculated. For this reason it is required through an on-site inspection, the features of the river cross section to be established, under which the maximum waters were flowed in the past. Since we have no data available from the on-site inspection, the maximal water quantity was determined by two methods – by analogy through regional empiric dependences and by the maximum rainfalls

2. Calculation method by empirical formulas

The maximum water quantities based on the available data within a certain region can be determined with dependences of the water quantity or of the runoff modulus from the surface area of the catchment basin, i.e., dependences:

$$\mathbf{Qmax = f\,(F)} \text{ or } \mathbf{Mmax = f\,(F)} \qquad \mathbf{(1)}$$

From the data available within the region for the hydro-metric points, the modulus of the maximum runoff for security rates of 0,1%, 1%, 2% and 5% were calculated.

The check for linear dependence existence showed that the determination coefficient is from 0.25 up to 0.45, the calculated values as compared to the data obtained from the hydro-metric point gives great deviations from the real values as measured at the point, which means that this dependence should not be used for calculation purposes.

The power dependence of the modulus of the maximum runoff and the respective security rate was considered:

$$M_{p\%} = A.F^{\square n} \qquad \mathbf{(1)}$$

where: M is the modulus of the runoff for the respective security rate

A – a coefficient

F – catchment surface area in sq. km

n – exponent

$p\%$ - security rate in %

The dependences as received and the calculated values from these dependences are shown in the following three figures, and in a generalized form the parameters A and n are given:

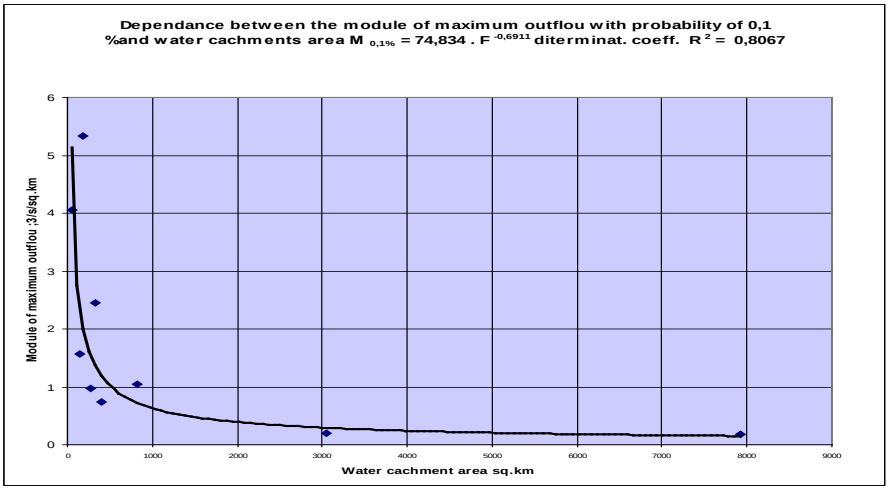


Figure. 1: Calculated dependency 0,1%

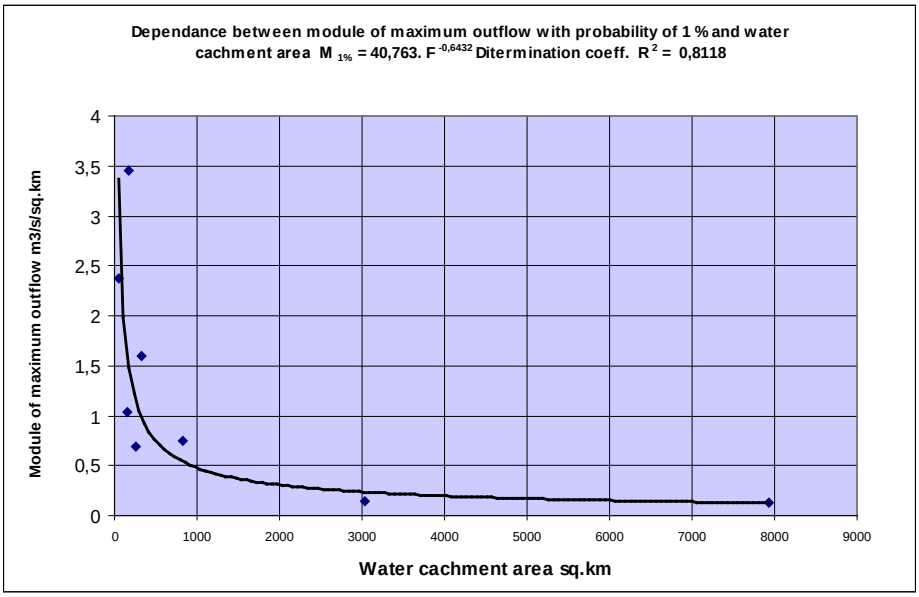


Figure. 2: Calculated dependency 1%

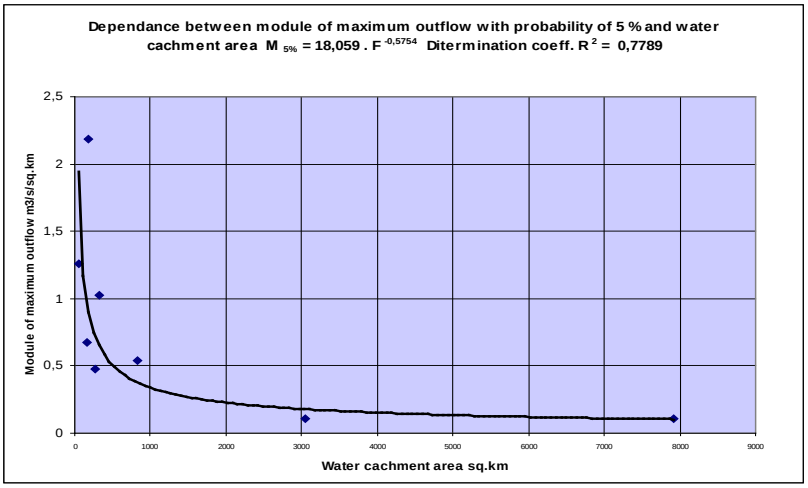


Figure. 3: Calculated dependency 5%

Parameter	Probability		
	0,1%	1%	5%
Empiric dependence			
Coefficient A	74,834	40,763	18,059
Coefficient n	-0,6911	-0,6432	-0,5754
Determination coefficient R2	0,8067	0,8118	0,7789

Table 7: The values of coefficients in the regional dependency for calculation of the „high" waves with different security rates

By the coefficients thus determined from dependence (2), the peaks of the high flow with different security rates for the bigger rivers (with a surface area more than 20 sq. km) in the region were determined.

No	Name	„High" waters with intensity measured in m3/s		
		0,1%	1%	5%
		m3/sec	m3/s	m3/s
1	2	3	4	5
1	Siva river	210,648	134,716	74,903
2	Mezeshka river	223,844	144,511	81,427
3	Golyama rekar	366,833	255,675	160,562
4	Levka river	347,480	240,160	149,035
5	Selskata reka	230,984	149,848	85,019
6	Kalamitza river	272,274	181,197	106,584

Table 8: The peaks of the „high" waves with different security rates for the bigger rivers being feeders of Maritza river in the section of Svilengrad municipality

With the dependencies we calculate also the high waves by maximal precipitations. The method gives very nice results compatible with HEC-RAS simulations.

3. Smart Water project tool

The state of the art for the flood hazard mapping for Svilengrad municipality has been based on the existing official maps that are published on both sites of Fire Fighting & Civil Protection Directorate of the Ministry of Interior and Basin Directorate, Ministry of Environment and Water in Bulgaria [3], [4]. These maps were developed on the basis of historical and prognosis flooding data in 2012 for the territory of whole Bulgaria. The goal of the Smart water tool is to use the collected data for the territory of Svilengrad municipality and by usage of the different dependencies formulas to estimate as accurate as possible the hydrological stage of the river Maritza in the vulnerable area of the Bulgarian-Turkish-Greek border zone.

3.1. Smart Water project tool structure

The project Smart Water has technical specifications which are oriented to the civil protection engineers, who could apply field response for the population in risk by having webGIS tool that could support their decision making in cases of large flood events. The test areas are river sections defined for each project partner and the Bulgarian region is on the

territory of municipality Svilengrad. The end user needs for the test cases cover the following types of information for the river monitoring:

- Distance from water level to river bank side
- Flooding areas
- Speed and direction of the water
- Water blades
- A series of maps of predefined and variable flood scenarios, with greater frequency for the selected test case area provided in an information layer (i.e. raster images) corresponding to the information required by the civil protection units, where the reliability of forecasts is the main focus.
- A set of data in the form of graphs, tables, or files for download will be also available for the identified critical levels.
- For each simulation and for each point, the maximum water height independently from the moment, when it is reached, will display immediate worst scenario situation possible from the given initial conditions.

The standard WMS interface will be applied for displaying the hydrological model outputs on the webGIS platform. The maps in raster format like JPEG or PNG will give opportunity for punctual queries for the users. The cartographic data will be provided in alphanumeric information related to the predetermined number of positions along the route of the monitored water course, deemed to be especially critical. The identification of the strategic locations and data supply will have geomorphologic and hydrodynamic sets, where will be included DEM (Digital Elevation Model) for the catchment basin, ortophoto images for better justification of land use, meteorological data for precipitations and additional climatic conditions, along with water level discharges and topology of the river levees for the simulated areas. On Fig. 4 is given the structure of the information flow that the webGIS platform will have implemented in its last version.

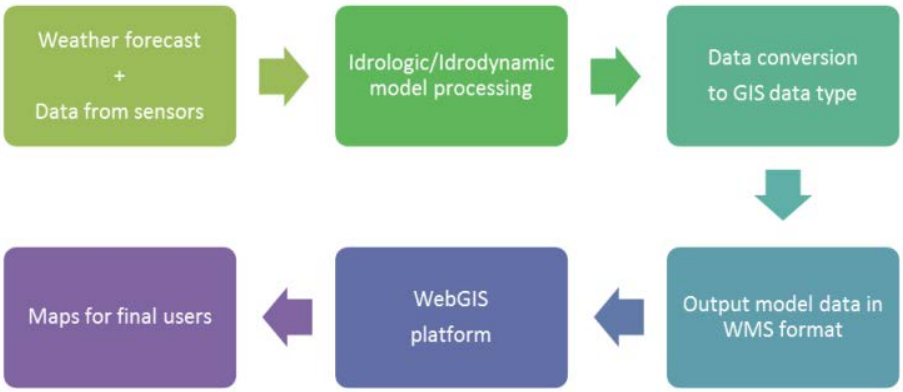


Figure. 4: Information flow as it will be implemented in the webGIS tool that will be the result of Smart Water project.

4. Conclusion

The presented work is still ongoing, because the project duration is until the end of January 2015. However the hydrological estimations for the vulnerable area of Svilengrad municipality are one of the first attempts of data collection and calculation as it is accepted according to the Bulgarian legislation based on INSPIRE directive and has priority to orient all its results to the webGIS tool, which will be of help in the everyday work of the Civil Protection engineers in the border area.

5. Acknowledgments

This paper has been supported by project Simple Management of Risk Through a Web Accessible Tool for EU Regions - ECHO/SUB/2012/638449. Acronym: SMART WATER. Web site: http://www.smartwaterproject.eu/.

References

[1] http://ec.europa.eu/environment/water/flood_risk/index.htm

[2] Gergov,G., 1971. Determination of the time of travel along the river network. In: Journal of Hydrology, Amsterdam. No 14, pp. 293-306.

[3] Republic of Bulgaria Flood Hazard Maps, Basin Directorate, Ministry of Environment and Water, Retrieved from: http://bd-ibr.org/details.php?p_id=0&id=243

[4] Republic of Bulgaria Civil Protection Flood hazard maps, Retrieved from: http://bsdi.asde-bg.org/data/Floods/GRZashtita/pdf/CE-BULGARIA-Flooded_Area_Districts_Land_Cover_Grazhdanska_zashtita.pdf

Residual bound of the matrix equations $X = A_1 + \sigma A_2^{\mathrm{H}} X^{-2} A_2$, $\sigma = \pm 1$

Ivan P. Popchev, Vera A. Angelova
Inst. of Inf. and Commun. Techn. - BAS
Akad. G. Bonchev, Str., Bl.2, Sofia, Bulgaria
e-mails: {ipopchev, vangelova}@iit.bas.bg

Abstract: Residual bound for the non-linear complex matrix equations $X = A_1 + \sigma A_2^{\mathrm{H}} X^{-2} A_2$, $\sigma = \pm 1$ is derived using the method of the Lyapunov majorants and the technique of the fixed point principles. The effectiveness of the bound is illustrated by a numerical example of order 5.

Keywords: perturbation analysis, residual bound, non-linear matrix equation

1 Introduction

We consider the non-linear complex matrix equations

$$X = A_1 + \sigma A_2^{\mathrm{H}} X^{-2} A_2, \quad \sigma = \pm 1, \tag{1}$$

where A_2 is a complex matrix and X, A_1 are Hermitian positive definite complex matrices.

The area of a practical application of equations (1) with $Q = I$ is discussed in [2, 3]. Studies of the necessary and sufficient conditions for the existence of Hermitian positive definite solutions in case $\sigma = +1$ and A_2 normal are given in [5]. Iterative algorithms for obtaining Hermitian positive definite spolutions are proposed in [2, 3, 5]. Perturbation bounds for the solutions are derived in [1].

In this paper a residual bound for the accuracy of the solution obtained by an iterative algorithm is derived. The bound is of a practical use as an effective measure for iterations termination.

Throughout the paper, the following notations are used: $\mathbb{C}^{n\times n}$ is the set of $n \times n$ complex matrices; A^{H} is the complex conjugate and $A^{\top}$ is the transpose of the matrix A; $A \otimes B = (a_{ij}B)$ is the Kronecker product of A and B; $\mathrm{vec}(A) = [a_1^{\top}, a_2^{\top}, \ldots, a_n^{\top}]^{\top}$ is the vector representation of the matrix A, where $A = [a_{ij}]$ and $a_1, a_2, \ldots, a_n \in \mathbb{C}^n$ are the columns of A; $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ are the spectral and the Frobenius matrix norms, respectively, $\|\cdot\|$ is a unitary invariant norm such as the spectral norm $\|\cdot\|_2$ or the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. The notation ':=' stands for 'equal by definition'.

The paper is organized as follows. The problem is stated in Section 2. In Section 3 a residual bound expressed in terms of the computed approximate solution to equations (1) is obtained using the method of Lyapunov majorants and the techniques of fixed point principles. In Section 4 the effectiveness of the bound proposed is demonstrated by a numerical example of 5th order.

2 Statement of the problem

Denote by $\hat{X} = X + \delta X$ the Hermitian positive definite solution of (1) obtained by some iterative algorithm. The obtained numerical solution $\hat{X}$ approximates the accurate solution X of (1), and the term δX, for which $\|\delta X\|_{\mathrm{F}} \le \varepsilon \|X\|_2$ is fulfilled, reflects the presence or round-off errors and errors of approximation in the computed with machine precision ε solution $\hat{X}$. Denote by

$$R(\hat{X}) := \hat{X} + \sigma A_2^{\mathrm{H}} \hat{X}^{-2} A_2 - A_1 \tag{2}$$

the residual of (1) with respect to $\hat{X}$.

The goal of our investigation is to estimate by norm the error δX in the obtained solution $\hat{X}$ of (1) in terms of the residual $R(\hat{X})$.

For this purpose applying the matrix inversion lemma

$$(X+\delta X)^{-1} = X^{-1} - \hat{X}^{-1}\delta X X^{-1} = (\hat{X}-\delta X)^{-1} - \hat{X}^{-1}\delta X X^{-1},$$

we rewrite equation (1) as an equivalent matrix equation

$$\delta X = R(\hat{X}) - \sigma A_2^{\mathrm{H}} X^{-2}\delta X \hat{X}^{-1} A_2 - \sigma A_2^{\mathrm{H}} X^{-1}\delta X \hat{X}^{-2} A_2, \tag{3}$$

or written in an operator form

$$\delta X = F(R(\hat{X}), \delta X), \tag{4}$$

where $F(S,H) : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is a linear operator, defined for some arbitrary given matrices W, V:

$$F(S,H) \;=\; S - \sigma W^{\mathrm{H}}(V-H)^{-2}HV^{-1}W - \sigma W^{\mathrm{H}}(V-H)^{-1}HV^{-2}W. \tag{5}$$

Taking the vec operation on both sides of (3) we obtain the vector equation

$$\begin{aligned} \mathrm{vec}(\delta X) &= \mathrm{vec}(F(R(\hat{X}),\delta X)) := \pi(\gamma, x) \\ \pi(\gamma,x) &= \gamma - \sigma(A_2^{\top}\hat{X}^{-1}\otimes A_2^{\mathrm{H}})\mathrm{vec}(X^{-2}\delta X) \\ &- \sigma(A_2^{\top}\hat{X}^{-2}\otimes A_2^{\mathrm{H}})\mathrm{vec}((\hat{X}-\delta X)^{-1}\delta X), \end{aligned} \tag{6}$$

where $\gamma := \mathrm{vec}(R(\hat{X}))$ and $x := \mathrm{vec}(\delta X)$. As in practice only the calculated approximate solution $\hat{X}$ is known, we represent in (6) the accurate solution X by the calculated approximate solution $\hat{X}$ and the error δX to be estimated: $X = \hat{X} - \delta X$.

3 Residual bound

Taking the spectral norm of both sides of (6), we obtain

$$\begin{aligned} \|\delta X\|_{\mathrm{F}} = \|\pi(\gamma,x)\|_2 &\le \|R(\hat{X})\|_{\mathrm{F}} + \|A_2^{\top}\hat{X}^{-1}\otimes A_2^{\mathrm{H}}\|_2\|X^{-2}\|_2\|\delta X\|_{\mathrm{F}} \\ &+ \|A_2^{\top}\hat{X}^{-2}\otimes A_2^{\mathrm{H}}\|_2\|(\hat{X}-\delta X)^{-1}\|_2\|\delta X\|_{\mathrm{F}}. \end{aligned} \tag{7}$$

To simplify the expression of the error δX in the obtained solution $\hat{X}$ and to avoid neglecting of higher order terms, we approximate $\|X^{-2}\|_2$ by $\|\hat{X}^{-1}\|_2\|(\hat{X}-\delta X)^{-1}\|_2$, admitting some rudeness in the bound.

Based on the nature of δX we can assume that $\|\delta X\|_{\mathrm{F}} \le \frac{1}{\|\hat{X}^{-1}\|_2}$. Then, it follows for $\|(\hat{X}-\delta X)^{-1}\|_2$ that

$$\|(\hat{X}-\delta X)^{-1}\|_2 \le \frac{\|\hat{X}^{-1}\|_2}{1-\|\hat{X}^{-1}\|_2\|\delta X\|_{\mathrm{F}}} \tag{8}$$

Replacing (8) in (7) we obtain

$$\begin{aligned} \|\delta X\|_{\mathrm{F}} = \|\pi(\gamma,x)\|_2 &\le \|R(\hat{X})\|_{\mathrm{F}} + \frac{\|A_2^{\top}\hat{X}^{-1}\otimes A_2^{\mathrm{H}}\|_2\|\hat{X}^{-1}\|_2^2}{1-\|\hat{X}^{-1}\|_2\|\delta X\|_{\mathrm{F}}}\|\delta X\|_{\mathrm{F}} \\ &+ \frac{\|A_2^{\top}\hat{X}^{-2}\otimes A_2^{\mathrm{H}}\|_2\|\hat{X}^{-1}\|_2}{1-\|\hat{X}^{-1}\|_2\|\delta X\|_{\mathrm{F}}}\|\delta X\|_{\mathrm{F}}. \end{aligned} \tag{9}$$

Denote by $\delta := \|\delta X\|_{\mathrm{F}}$, $r := \|R(\hat{X})\|_{\mathrm{F}}$, $\chi := \|\hat{X}^{-1}\|_2$, $\alpha_1 := \|A_2^{\top}\hat{X}^{-1} \otimes A_2^{\mathrm{H}}\|_2\|\hat{X}^{-1}\|_2^2$, $\alpha_2 := \|A_2^{\top}\hat{X}^{-2} \otimes A_2^{\mathrm{H}}\|_2\|\hat{X}^{-1}\|_2$. For equation (9) we obtain

$$\begin{aligned} \delta &\leq r + \frac{\alpha_1\delta}{1-\chi\delta} + \frac{\alpha_2\delta}{1-\chi\delta} \\ &\leq r + a_1\delta + a_2\delta^2, \end{aligned} \tag{10}$$

with $a_1 := \alpha_1 + \alpha_2 - \chi r$, $a_2 := \chi$.

To estimate the norm of the operator $F(R(\hat{X}), \delta X)$ we apply the method of Lyapunov majorants. We construct a Lyapunov majorant equation with the quadratic function $h(r, \rho)$

$$\rho = h(r, \rho), \quad h(r, \rho) := r + a_1\rho + a_2\rho.$$

Consider the domain

$$\Omega = \{r : a_1 + 2\sqrt{ra_2} \leq 1\}. \tag{11}$$

If $r \in \Omega$ then the majorant equation $\rho = h(r, \rho)$ has a root

$$\rho = f(r) := \frac{2r}{1 - a_1 + \sqrt{(1-a_1)^2 - 4ra_2}}. \tag{12}$$

Hence, for $r \in \Omega$ the operator $\pi(r, .)$ maps the closed convex set $\mathcal{B}_{f(r)} \subset \mathbb{R}^{n^2}$ into itself. The set $\mathcal{B}$ is small, of diameter $f(r)$ and $f(0) = 0$. Then, according to the Schauder fixed point principle, there exists a solution $\xi \in \mathcal{B}_{f(r)}$ of (4) and hence $\|\delta X\|_{\mathrm{F}} = \|\xi\|_2 \leq f(r)$. In what follows, we deduced the following statement.

Theorem 1. Consider equations (1) for which the solution X is approximated by $\hat{X}$, obtained by some iterative algorithm with residual $R(\hat{X})$ (2).

Let $r := \|R(\hat{X})\|_{\mathrm{F}}$, $\alpha_1 := \|A_2^{\top}\hat{X}^{-1} \otimes A_2^{\mathrm{H}}\|_2\|\hat{X}^{-1}\|_2^2$, $\alpha_2 := \|A_2^{\top}\hat{X}^{-2} \otimes A_2^{\mathrm{H}}\|_2\|\hat{X}^{-1}\|_2$ and $\chi := \|\hat{X}^{-1}\|_2$.

For $r \in \Omega$, given in (11) the following bounds are valid:

- non-local residual bound

$$\|\delta X\|_{\mathrm{F}} \leq f(r), \quad f(r) := \frac{2r}{1 - a_1 + \sqrt{(1-a_1)^2 - 4ra_2}}, \tag{13}$$

where $a_1 := \alpha_1 + \alpha_2 - \chi r$, $a_2 := \chi$;

- relative error bound in terms of the unperturbed solution X

$$\frac{\|\delta X\|_{\mathrm{F}}}{\|X\|_2} \leq \frac{f(r)}{\|X\|_2}. \tag{14}$$

- relative error bound in terms of the computed approximate solution $\hat{X}$

$$\frac{\|\delta X\|_{\mathrm{F}}}{\|X\|_2} \leq \frac{f(r)/\|\hat{X}\|_2}{1 - f(r)/\|\hat{X}\|_2}. \tag{15}$$

4 Experimental results

To illustrate the effectiveness of the bound, proposed in Section 3, we construct a numerical example on the base of Example 4.3. from [4]. Consider equation $X + A_2^{\mathrm{H}} X^{-2} A_2 = A_1$ with coefficient matrices

$$A_2 = \frac{1}{10}\begin{pmatrix} -1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix}, \quad A_1 = X + A_2^{\mathrm{H}} X^{-2} A_2$$

and solution $X = \mathrm{diag}(1,\ 2,\ 3,\ 2,\ 1)$. The approximate solution $\hat{X}$ of X is chosen as

$$\hat{X} = X + 10^{-2j} X_0; \quad X_0 = \frac{1}{\|C^\top + C\|}(C^\top + C),$$

where C is a random matrix, generated by MatLab function **rand**. The norm of the relative error $\|\delta X\|_{\mathrm{F}}/\|X\|_2$ in the computed solution $\hat{X}$ is estimated with the relative error bound (15) for $\hat{X}$, defined in Theorem 1

The results for $j = 1, 2, 3, 4, 5$ are listed in Table 1.

Table 1.

j	1	2	3	4	5
$\frac{\|\delta X\|_{\mathrm{F}}}{\|X\|_2}$	3.33×10^{-3}	3.33×10^{-5}	3.33×10^{-7}	3.33×10^{-9}	3.33×10^{-11}
est (15)	4.19×10^{-3}	4.17×10^{-5}	4.17×10^{-7}	4.17×10^{-9}	4.17×10^{-11}

The results show that the residual bound proposed in Theorem 1 is quite sharp and accurate.

Acknowledgments

The research work presented in this paper is partially supported by the FP7 grant AComIn No 316087, funded by the European Commission in Capacity Programme in 2012-2016.

References

[1] Angelova V.A. (2003) Perturbation analysis for the matrix equation $X = A_1 + \sigma A_2^H X^{-2} A_2$, $\sigma = \pm 1$. *Ann. Inst. Arch Genie Civil Geod., fasc II Math.*, 41, 33–41.

[2] Ivanov I.G., El-Sayed S.M. (1998) Properties of positive definite solutions of the equation $X + A^* X^{-2} A = I$. *Linear Algebra Appl.*, 297, 303–316.

[3] Ivanov I.G., Hasanov V., Minchev B. (2001) On matrix equations $X \pm A^* X^{-2} A = I$. *Linear Algebra Appl.*, 326, 27–44.

[4] Xu S. (2001) Perturbation analysis of the maximal solution of the matrix equation $X + A^* X^{-1} A = P$. *Linear Algebra Appl.*, 336, 61–70.

[5] Zhang Yuhai (2003) On Hermitian positive definite solutions of matrix equation $X + A^* X^{-2} A = I$. *Linear Algebra Appl.*, 372, 295–348.

Scalable System for Financial Option Prices Estimation

D. Dimitrov and E.Atanassov
Institute of Information and Communication Technologies
Acad. G. Bonchev str. 25A, Sofia, Bulgaria
d.slavov@bas.bg , emanouil@parallel.bas.bg

Abstract: In this paper we describe a production-ready ESB system for estimation of option prices using stochastic volatility models. The Heston model is used as a basis for modelling the evolution of asset prices. Our framework allows for accurate calibration of the Heston model to the market data, using GPGPU-computing. The Zato framework is used as an integration layer, while the main computations are distributed to HPC resources. In the first section we present the motivation for our research and the main building blocks for our system. In the next section we discuss our approach to calibration of the Heston model. In Section 3 we present the whole setup of the system. In Section 4 we give examples of the operation of the system and show performance data. The system is capable of using different data sources and is extendable from both infrastructure and software point of view with hot deployment. The main advantage of the system is that by incorporating GPGPU-computing nodes it allows for the use of more accurate models that are otherwise unfeasible.
Keywords: Option pricing, Heston model, GPGPU.

1. Introduction

A financial option is a contract which gives to the owner the right, but not the obligation, to buy or sell an underlying asset or instrument at a specified strike price on or before a specified date. Valuation of options is one of the most important problems in the world of Financial Mathematics. There are many approaches to this problem, including for example Monte Carlo[1] simulation of the evolution of the price of a financial asset or computations based on Fourier Transforms[2]. In this paper we build upon our previous work on the Heston model[3] in order to incorporate the estimation of values of financial options in a production ready high performance system.

The Heston model[4] is one of the most popular stochastic volatility models for estimation of financial derivatives. The model was proposed by Steven Heston in 1993 and takes into account non-lognormal distribution of the assets returns, the leverage effect and the important mean-reverting property of volatility:

$$dX(t)/X(t) = rdt + \sqrt{V(t)}dW_X(t)$$

$$dV(t) = k(\theta - V(t))dt + \varepsilon\sqrt{V(t)}dW_V(t)$$

$X(t)$ is asset price process, k, θ, ε are constants, $V(t)$ is instantaneous variance and W_X, W_V- Brownian motions. The initial conditions are $X(0) = X_0$ and $V(0) = V_0$ .We assume that $< dW_X(t), dW_X(t) >= \rho dt$ where □ is correlation parameter. There are many ways to discretize and simulate the model but one of the most widely used is the Monte Carlo one, where one discretizes along the time and simulates the evolution of the price of the underlying. For discretization scheme we use Andersen[5] which sacrifices the full unbiasedness, achieved under the exact scheme of Broadie and Kaya [6], to attain much faster execution with similar accuracy. It is known that Monte Carlo simulations are computationally intensive that is why we have developed a GPGPU algorithms [7] to achieve fast execution times. The General Purpose GPU computing uses graphic cards as co-processors to achieve powerful and cost efficient computations. The higher class devices have large number of transistors and hundreds to thousands of computational cores which makes them efficient for Monte Carlo simulations because there is a large degree of separate, independent numerical trajectories with low amount of synchronization between them. In our work we use NVIDIA graphic cards with their parallel computing architecture CUDA[8].

In order to achieve a production ready system that computes option prices in a near real time manner and can be dynamically scaled in heterogeneous environment we used Zato framework[9] as base integration system and Quandl[10] as main resource of financial data.

Zato is, an open-source ESB (Enterprise Service Bus) middleware and backend server written in Python, designed to provide easy lightweight integration for different systems and services. The platform does not have any restrictions for the architecture and can be used to provide SOA (Service Oriented Architecture). The framework supports out of the box HTTP, JSON, SOAP, SQL, Messaging (AMQP, JMS WebSphere MQ, ZeroMQ), NoSQL, FTP. It can be managed via browser-based admin UI, CLI, API and provides security, statistics, job scheduling, load-balancing and hot-deployment.

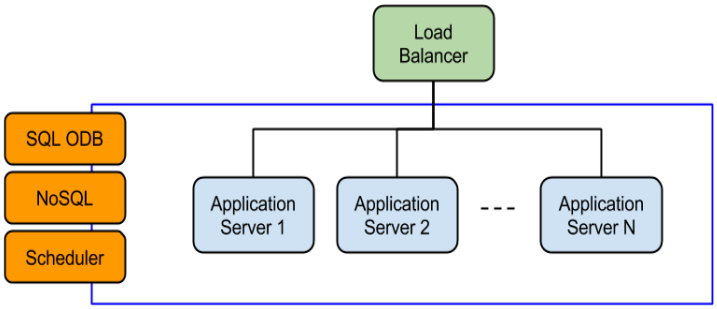


Figure 1 Zato components

It has several components that build the whole environment:

- Load Balancer - implemented using HAProxy can handle large number of incoming connections.
- Application servers - based on gunicorn which allows them to be asynchronous, fast, light on resources and handle large numbers of incoming HTTP connections. When the environment is set up there are several (minimum 2) servers one of which is pointed as singleton server. This role makes him responsible for managing tasks that should not be executed in different instances - job's scheduler and messaging connectors (AMQP, JMS, ZeroMQ). Each server in the cluster always has the same set of services as the others and is always active.
- SQL ODB - database for storing Zato configurations, statistical information and other user defined data. It supports Oracle and PostgreSQL.
- NoSQL - a Redis server used for messaging broker between the servers and scheduler and servers. Services and connectors communicate asynchronous, indirectly through Redis.
- Scheduler - is used for periodically invoked services and supports the following type of jobs: one-time, interval-based and cron-style.

Quandl is a cloud big-data platform/search engine for various financial, economic and social datasets like FX rates, futures, commodities, interest rates, stock prices, business indexes and economic indicators. All data is available via an API and there are a lot libraries that wrap it in various languages - C/C++, Java, Python, R, Matlab and etc. The platform contains more than 10 million datasets.

In the next section we explain our approach towards the calibration of the Heston model based on incoming financial data. The third section describes the setup of our

production-ready system and the interconnection of its various parts. After that we outline our observations and experience from testing the system and we finish with conclusions and directions for future work.

2. Calibration of the Heston model

Although the prices of financial assets are apparently unpredictable, they exhibit some features and are amenable to the use of the mathematical apparatus of stochastic processes. The main building blocks for describing the dynamics of the price of a financial asset or a system of correlated financial assets are the Brownian motion and the Poisson process. The volatility of the price can be modeled as either deterministic or stochastic process. The stochastic volatility models are more powerful, but are more demanding in terms of computational power. Once we decide on the model to be used, we have to solve the problem of calibration of the parameters of the model, using the observed data. This task becomes more computationally intensive and mathematically challenging with the increased number of degrees of freedom of the model. In many practical situations it has to be accomplished within limited time and computational resources. The classical model of Black-Scholes[11] is still used today, although it assumes log-normal distribution of the asset returns. It has been found to be insufficient to account for the specifics of market behaviour and accomplish acceptable pricing of traded options, but is still used as a benchmarking tool. For example, the implied volatility, obtain through the Black-Scholes formula, can be used as a way of measuring the perceived risk. The Heston model is a more complex model which assumes the instantaneous stochastic volatility follows an Ornstein–Uhlenbeck process. Unfortunately, the instantaneous stochastic volatility is not directly observable.

One of the approaches for calibration of the parameters of the model, that we have employed, is based on the fitting of the so-called volatility surface[12]. On the next figure 2 one can see the implied volatility surface, computed for some parameters of the Heston model. The financial options that are traded on the market produce data points that deviate from this surface. For each observed option price, one obtains one data point. Thus the usual approach for calibration is based upon the fitting of these data points with the implied volatility surface, by defining a suitable loss function (usually mean-squared error) and optimizing over the set of all admissible parameters.

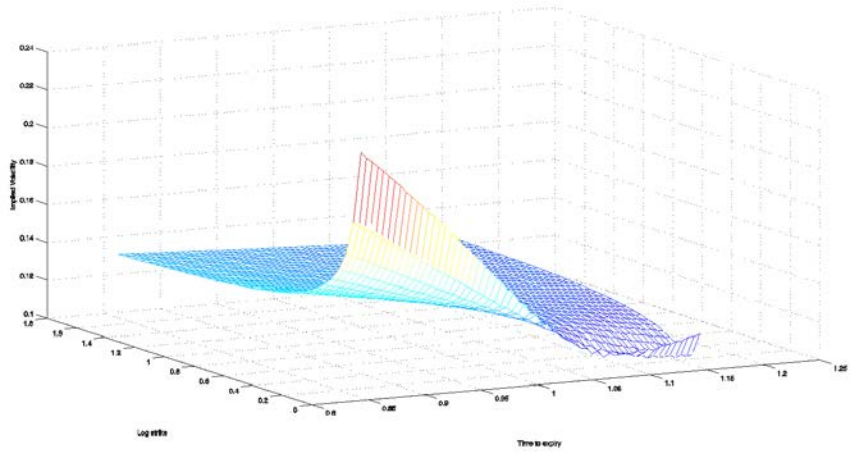


Figure 2. Implied volatility as a function of normalized strike price and time to expiry. Resulting from Heston model fit of SPY option prices

This approach has certain drawbacks. First of all, it results in rapid changes in the parameters of the model as the market situation changes. It is also dependent on the global optimization routine being used, since many such routines end-up in local instead of global minima. In many cases the final values of the objective (loss) function obtained under different optimization routines, are close, but the values of the model parameters are different.

That is why we decided to make use of more computational resources and to add more terms in the loss function, in order to obtain more stable results. Our loss function is based not only on the current market prices, but also on the prices in the recent past, e.g., if daily observations are considered, the loss function can be defined as:

$$\sum_{t=0}^{k-1} e^{-ct} \sum_{j=1}^{n} \boldsymbol{w_j(t)}\left(\sigma_j(\boldsymbol{t}) - \sigma_j(\boldsymbol{t}, \boldsymbol{\Pi})\right)^2$$

where k is the total number of time periods used, c is a constant, $\boldsymbol{w_j}$ is the weight**,** $\sigma_j(t)$ is the observed price at time t of the option j, while $\sigma_j(\boldsymbol{t}, \boldsymbol{\Pi})$ is the computed value of the option $\boldsymbol{j}$ under the set of model parameters $\boldsymbol{\Pi}$**.**

This approach is generic and applies not only to the Heston model, but in our system we implemented an algorithm for fast computation of the option prices under the Heston model that is suitable for using GPGPU resources - the so-called COS algorithm. Several routines for global minimization were tested, mainly from the package for non-linear optimization that is called NLOPT[13][14]. The weights can be defined in many ways to take

into account the amount of uncertainty in the observed prices. In our tests we defined the weight to be equal to 3 if both open interest and volume for the particular option are non-zero in the given day, 2 if one of these is zero and the other is non-zero, and 1 if the price is only result from bid-ask quotes.

This type of calibration procedure depends on availability of pricing data for the options. In many cases such data is hard to obtain or not relevant. For example, for very short term options, e.g., minutes, the data about options with duration approximately one month is not suitable. For such cases one has to develop approaches based on using only data about prices of the assets. There is large amount of theory and practical implementations that deal with such cases. In our system we used the Andersen scheme as an approximation and combining it with the Markov-Chain Monte Carlo approach, in order to produce chains that sample sets of Heston model parameters. This scheme was also implemented using parallelization across multiple GPGPUs. The parameters that are obtained in this way may be different from those that the above algorithm provides, because they use different sets of data and give more weight to data points coming from relatively distant times. It is advisable to use the former approach when option data is available and judged to be relevant, while the latter will be used as fallback.

3. System setup

The Zato framework is installed with two application servers, load-balancer, SQL database (PostgreSQL) and Redis in our datacenter. For host machines we used virtual ones because, the asynchronous nature of the platform it is designed for more data intensive operations so virtual machines offer a good flexibility and easy scalability. The front load-balancer is HA-proxy which is a fast and reliable solution for high availability, load balancing and proxying for TCP and HTTP. The data integration is done via services that consume the Quandl public API and synchronize the data with the internal SQL database. For ease and real-time communication we have set a Twitter account and connected it with the integration backend services via the Twitter API. We have used the AMQP solution provided by the platform to integrate the internal HPC/GPU resources with Zato. AMQP stands for Advanced Message Queuing Protocol an open standard application layer protocol for message oriented middleware. These resources include a cluster based on HP Cluster Platform Express 7000 with 36 identical blades of the type HP BL 280c, equipped with dual Intel Xeon X5560 @ 2.8GHz and 24 GB RAM per blade and a separate configuration for the GPUs - two HP ProLiant SL390s G7 servers with Intel(R) Xeon(R) CPU E5649 @

2.53GHz, which are equipped with 8 NVIDIA Tesla M2090 cards each. A card of that model has 512 cores for GPU computing and achieves 1331 GFLOPS in single precision and 665 GFLOPS in double precision. All code for the GPUs is written on CUDA - NVIDIA parallel computing platform for graphic cards. All servers are interconnected with non-blocking InfiniBand interconnection, thus keeping all servers under the same connectivity parameters. The whole Zato configuration can be modified easily via command line or web admin interface. The same is valid for the services and can be dynamically changed on the fly. All services have REST interface which offers easy use and setup of clients.

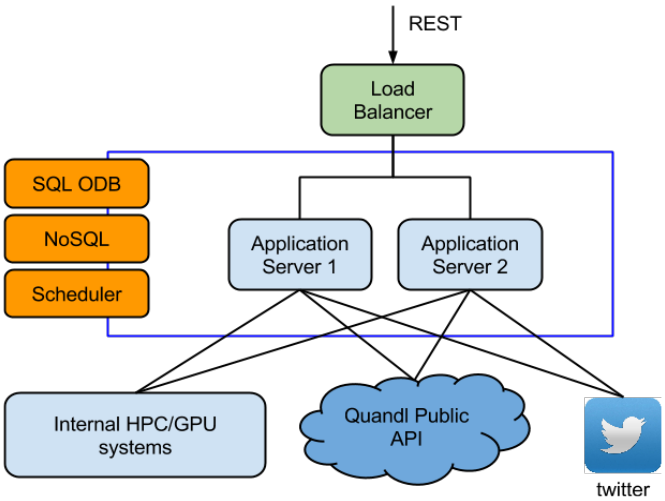


Figure 3 System components

4. Operations and performance

The system is configured to update the financial data on daily basis and persist it in PostgreSQL database. The logical organization offers easy horizontal scalability if more datasets are needed. At this moment we consume only the Quandl Financial Data API but it wraps Google Finance, Yahoo Finance, NASDAQ and 240 more financial data providers with more than 230 000 datasets with financial assets history. In case of need the system can be easily updated with another provider by API or static import of data in the database. Quandl API is flexible and allows incremental update of the targeted datasets which is done by the scheduler of the framework. In addition we have connected it with Twitter in order to have the ability to config real-time monitoring and notification service. The current setup is with two applications servers which is the minimal configuration but the application server layer can be easily scaled up and down via the admin interfaces. Also the framework offers dynamic deployment and hot replacement of the services without any restarts or downtimes.

This comes from the fact that it is Python framework and services are compiled runtime. The message-passing communication between the data layer and the computational resources provides asynchronous delivery of the data and fail over solution in case of broken connection between the sender and receiver. The table below shows the calibration algorithm running times on NVIDIA Tesla m2090 cards with calculated parallel efficiency.

GPU	1	2	3	4	5	6	7	8
time(s)	308	158	112	90	79	69	66	63
efficiency	1.0	0.98	0.92	0.85	0.78	0.74	0.66	0.61

Figure 4 Parallel efficiency for 100000 iterations, n=18 (days) and 34938 options

As it is shown, the running times range from one up to several minutes which offers a good near real-time processing and simulation of the prices.

5. Conclusions

The presented setup and results show a scalable system for real-time option pricing. As it was presented -every layer of the architecture is dynamic. On data layer, the organization of the datasets offers easy management and clusterization if needed. The business logic layer – Zato application servers are in cluster mode by default. These servers use messaging to communicate with the infrastructure layer – the HPC/GPU resources - an approach that provides ease extension if more computational resources are needed. The developed GPGPU algorithms offers time-efficient parameter estimation and simulation of Heston stochastic volatility model for option pricing. For future work we intend to implement and add more models in order to have a system with rich portfolio for financial technical analysis.

Acknowledgment

This research has been partially supported by the National Science Fund of Bulgaria under Grants DCVP 02/1 (SuperCA++) and the European Commission under EU FP7 project EGI-InSPIRE (contract number RI-261323).

References

[1] P. Glasserman, Monte Carlo Methods in Financial Engineering, Springer, New York, 2003

[2] Fang, F. and C. W. Oosterlee, A novel pricing method for European Options based on Fourier-Cosine Series Expansions, SIAM Journal on Scientific Computing 31(2), (2008), pp 826--848.

[3] E. Atanassov, D. Dimitrov and S. Ivanovska, Efficient Implementation of the Heston Model Using GPGPU, Monte Carlo Methods and Applications, De Gruyter, 2012, pp. 21-28, ISBN: 978-3-11-029358-6, ISSN: 0929-9629 .

[4] S. Heston, A closed-form solution for options with stochastic volatility, Review of Financial Studies, 6, (1993), 327--343.

[5] L. B. G. Andersen, Efficient Simulation of the Heston Stochastic Volatility Model, Banc of America Securities, 2007, available online at http://ssrn.com/abstract=946405 (last accessed September 21, 2012).

[6] M. Broadie and Ö. Kaya, Exact simulation of stochastic volatility and other affine jump diffusion models, Operation Research 54 (2006), 217–231.

[7] E. Atanassov and S. Ivanovska, Sensitivity Study of Heston Stochastic Volatility Model Using GPGPU, LSSC 2011, LNCS, Springer, 2012, Volume 7116/2012, pp. 439-446, 2012, DOI: 10.1007/978-3-642-29843-1_49, ISSN: 0302-9743.

[8] CUDA, http://developer.nvidia.com/category/zone/cuda-zone

[9] Zato framework, https://zato.io/docs/index.html

[10] Quandl, https://www.quandl.com/about/data

[11] Black F. M. Scholes, "The pricing of options and corporate liabilities",Journal of Political Economy, 81(3), 1973.

[12] Jim Gatheral, The Volatility Surface: A Practitioner's Guide}, Wiley Finance, 2006

[13] Steven G. Johnson, The NLopt nonlinear-optimization package, http://ab-initio.mit.edu/nlopt

[14] M. Galassi et al, GNU Scientific Library Reference Manual (3rdEd.), ISBN 0954612078, http://www.gnu.org/software/gsl/

Preferences and Modeling in Mathematical Economics: Utility Approach

Yuri P. Pavlov

Bulgarian Academy of Sciences, Institute of Information and Communication Technologies

Sofia, Bulgaria

e-mail: yupavlov15@isdip.bas.bg

Abstract: In the paper we demonstrate a system engineering value driven approach within determination of the optimal portfolio allocation modeled with Black-Scholes stochastic differential equation dynamic and determination of the equilibrium points in a competitive trading modeled by the Edgworth box. The solutions are specified on the individual consumers' preferences presented as utilty functions in sense of von Neuman. The mathematical formulations presented here could serve as basis of tools development. These value evaluation leads to the development of preferences-based decision support in machine learning environments and control design in complex problems.

Keywords: Preferences, Utility function, optimal portfolio, Hamilton-Jakoby-Belman equation, Edgeworth box, equilibrium

1 Introduction

Mathematical economics is a discipline that utilizes mathematical principles to create models to predict economic activity and to conduct quantitative tests. Although the discipline is heavily influenced by the bias of the researcher, mathematics allows economists to explain observable phenomenon and to permit theoretical interpretation. His fields are applications of mathematical techniques to the analyses of the theoretical problems and development of meaningful quantitative models for economic analyses. In these fields are included static equilibrium analyses, comparative statics (as to change from equilibrium to another by a change of important factors) and dynamic analysis tracing changes in an economic system over time. In the 19th century the economic modeling started the use of differential calculus to explain human behavior (preferences and objectives) and to describe analytically economic processes with human participations. In contemporary economic theory risk and uncertainty in the human decisions have been recognized as central topics. Along this line

Utility theory and the more general Measurement theory permit development of complex models in which human participation is reflected analytically [1].

The main assumption in each management or control human decision is that the values of the subject making the decision (DM) are guiding force and as such they are the main moment in supporting the decisions [5]. The utility theory is one of the approaches to measurement and utilization of such conceptual information and permits the inclusion of the decision maker (or the technologist) in the complex model „DM – quantitative model" in mathematical terms. Utility function based on DM's preferences as objective function allows for quantitative analysis in risk and removal of logical inconsistencies and errors which appear as uncertainty in economic solutions.

The main focus of the paper is the synchronous merger of mathematical exactness with the empirical uncertainty in the human notions in quantitative models in economic field [3]. The expression of the human notions and preferences contain characteristic of uncertainty due to the cardinal type of the empirical DM's information. The appearance of this uncertainty has subjective and probability nature [3, 4, 5, 8, 9]. Possible approach for solution of these problems is the stochastic programming. The uncertainty of the subjective preferences could be taken as an additive noise that could be eliminated as typical for the stochastic approximation procedures and stochastic machine-learning [12].

In the paper is proposed a methodology that is useful for dealing with the uncertainty of human behavior in complex problems. This methodology permits mathematical description of complex system "DM-economic process". We illustrate this methodology in two examples, two quantitative models for economic analyses and predictions. The first is „Optimal portfolio allocation" based on a utility objective function and Black-Scholes model dynamic as stochastic process for control [11]. The second is the well-known Edgeworth Box model in microeconomic [3]. In this example we demonstrate the approach by description of the consumer's demand's curves as utility function based on DM's preferences and determination of the contract curve in agreement with the DM's preferences.

The described approach permits representation of the individual's preferences as utility function, evaluated as machine learning and inclusion of this function in the quantitative economic model as objective function [1, 3].

2 Quantitative Modeling, Preferences and Utility

Here we describe applications of the proposed mathematical approach for representation of the individual DM's preferences as analytical utility function and their use

for model descriptions and solutions of economic problems with decisive human participations. In the first subsection we give a brief description of the approach for polynomial utility function approximation based on DM's preferences. In the following subsections we discuss two quantitative mathematical economic models. First is in the field of dynamic analysis with construction of optimal portfolio control and tracing changes in agreement with the DM's preferences. The second model is in the field of comparative statics for predictions the equilibrium in agreement with DM's preferences, the famous Edgeworth box model.

2.1 Ordering, Preferences and Utility Polynomial Approximation

We seek analytical representation of the individual's preferences as utility function, evaluated as recurrent stochastic programming and machine learning for value based decision making in the framework of the axiomatic decision making.

Measurement is an operation in which a given state of the observed object is mapped to a given denotation [7]. In the case when the observed property allows us not only to distinguish between states but to compare them by preference we use a stronger scale, the ordering scale. The preference relation in the ordering scale (x is preferable to y) is denoted by ($x \succ y$). If there exist incomparable alternatives, then the scale is called a scale of *partial ordering*. A "value" function is a function u(.) for which it is fulfilled [4, 5]:

$$((x, y) \in X^2, x \succ y) \Leftrightarrow (u(x) > u(y)).$$

Under this scale we cannot talk about distance between the different alternatives. Here only ordinal evaluations within different mathematical processing of the information may be used. If with the ordering of the alternatives we can evaluate the distance between them we can talk about interval scale [7, 12]. For these scales the distances between the alternatives have the meaning of real numbers. The transition from one interval scale to another is achieved with affine transformation: $x = ay + b, (x, y) \in X^2, a > 0, b \in R.$

Among these type of scales is also the measurement of the utility function through the so called "lottery approach". Once more we emphasize that here the calculations are done with numbers related to the distances between the alternatives, and not with the numbers relating to the alternatives themselves. For instance, if we say that a body is twice as warm as another in Celsius, this will not be true if the measurements were in Kelvin. Let $\mathbf{X}$ be the set of alternatives ($\mathbf{X} \subseteq \mathbf{R}^m$). The DM's preferences over $\mathbf{X}$ are expressed by ($\succ$). Let $\mathbf{A}_{u^*}$ and $\mathbf{B}_{u^*}$ are the sets: $\mathbf{A}_{u^*}=\{(x,y) \in \mathbf{R}^{2m} / (u^*(x)) > u^*(y)\}$, $\mathbf{B}_{u^*}=\{(x, y) \in \mathbf{R}^{2m} / (u^*(x)) < u^*(y)\}$.

If there is a function $\mathbf{F}(x,y)$ of the form $\mathbf{F}(x,y)=\mathbf{f}(x)-\mathbf{f}(y)$, positive over $\mathbf{A}_{u^*}$ and negative over $\mathbf{B}_{u^*}$, then the function $\mathbf{f}(x)$ is a value function equivalent to the empirical DM's value function u*(.). The construction of such functions of two variables is one of the ways for evaluating the value functions in ordinal aspect [12]. Such approach also permits the use of stochastic recurrent techniques for "pattern recognition" for solving the problems. In the deterministic case it is true that $\mathbf{A}_{u^*}\cap\mathbf{B}_{u^*}=\varnothing$. In the probabilistic case it is true that $\mathbf{A}_{u^*}\cap\mathbf{B}_{u^*}\neq\varnothing$ [ ]. Let $\boldsymbol{X}$ be the set of alternatives and $\boldsymbol{P}$ is a set of probability distributions over $\boldsymbol{X}$. A utility function $u(.)$ will be any function for which the following is fulfilled:

$$(p \succ q \ , (p,q) \in \boldsymbol{P}^2) \Leftrightarrow (\int u(.)dp > \int u(.)dq).$$

In keeping with Von Neumann and Morgenstern [4] the interpretation of the above formula is that the integral of the utility function $u(.)$ is a measure concerning the comparison of the probability distributions p and q defined over $\boldsymbol{X}$. The notation ($\succ$) expresses the preferences of DM over $\boldsymbol{P}$ including those over $\boldsymbol{X}$ ($\boldsymbol{X}\subseteq\boldsymbol{P}$). There are different systems of mathematical axioms that give sufficient conditions of a utility function existence. The most famous of them is the system of Von Neumann and Morgenstern's axioms [4]. The following proposition is almost obvious [4, 7].

Proposition1. The DM "*preference*" relation ($\succ$) defined by the utility function $u(.)$ is "*negatively transitive*" ($\neg(p \succ t) \wedge \neg (t \succ q)$) $\Rightarrow \neg(p \succ q)$).

The following proposition discusses the transitivity of the equivalence relation ($\approx$). This property is violated in practice most often times.

Proposition2. In the case of a "*negative transitivity*" of ($\succ$) the "*indifference*" relation ($\approx$) is transitive ($((x \approx y) \wedge (y \approx t)) \Leftrightarrow (x \approx t)$)).

Corollary: Let the preference relation ($\succ$) is "*irreflexive*" ($\neg(p \succ p)$) and "*negatively transitive*", then the "*indifference*" relation ($\approx$) is an "*equivalence*" (reflexive, symmetric and transitive).

The irreflexivity of the preferences and the negative transitivity of the preference relation split the set $\boldsymbol{X}$ into non-crossing equivalence classes. The factorized set of these classes is marked by $\boldsymbol{X}/\approx$. We need the next two definitions. A "*weak order*" is an asymmetric and "negatively transitive" relation. The transitivity of the "*weak order*" ($\succ$) follows from the "asymmetry" and the "negative transitivity". A "*strong order*" is a "*weak order*" for which is fulfilled ($\neg(x \approx y) \Rightarrow ((x \succ y) \vee (x \prec y))$. It is proven in [4] that the existence of a "*weak order*" ($\succ$) over $\boldsymbol{X}$ leads to the existence of a "*strong order*" preference relation ($\succ$) over $\boldsymbol{X}/\approx$. *Consequently the presumption of existence of a utility function u(.) leads to the*

existence of: asymmetry $(x \nmid y) \Rightarrow (\neg(x \nmid y))$, *transitivity* $(x \nmid y) \wedge (y \nmid z) \Rightarrow (x \nmid z)$, *and transitivity of the "indifference" relation* $(\approx)$.

So far we are in the preference scale. The assumption of equivalence with precision up to affine transformation has not been included. In other words we have only a value function. For value, however, the mathematical expectation is unfeasible, but we underline that the mathematical expectation is included in the definition of the utility function. In practice the utility function is measured by the lottery approach [5, 8]. There are quite different utility evaluation methods that based prevailing on the "lottery" approach (gambling approach). A "lottery" is called every discrete probability distribution over X. We mark as $<x,y,\alpha>$ the lottery: α is the probability of the appearance of the alternative *x* and $(1-\alpha)$ - the probability of the alternative *y*. The most used evaluation approach is the following assessment: $z \approx <x,y,\alpha>$, where $(x,y,z) \in \boldsymbol{X}^3$, $(x \nmid z \nmid y)$ and $\alpha \in [0,1]$. Weak points of this approach are violations of the transitivity of the relations and the so called "certainty effect" and "probability distortion" [12]. Additionally, the determination of alternatives *x* (*the best*) and *y* (*the worst*) on condition that $(x \nmid z \nmid y)$ where z is the analyzed alternative is not easy. Therefore, the problem of utility function evaluation on the grounds of expert preferences is a topical one. The violations of the transitivity of the relation $(\approx)$ also leads to declinations in the utility assessment. All these difficulties could explain the DM behavior observed in the famous Allais Paradox that arises from the "*independence*" axiom.

The determination of a measurement scale of the utility function *u*(.) originates from the previous mathematical formulation of the relations $(\nmid)$ and $(\approx)$. It is accepted that $(\boldsymbol{X} \subseteq \boldsymbol{P})$ and that $\boldsymbol{P}$ is a convex set $((q,p) \in \boldsymbol{P}^2 \Rightarrow (\alpha q + (1-\alpha)p) \in \boldsymbol{P}$, for $\forall \alpha \in [0,1])$. Then the utility function *u*(.) over $\boldsymbol{X}$ is determined with the accuracy of an affine transformation [4]:

Proposition3. If $((x \in X \wedge p(x)=1) \Rightarrow p \in \boldsymbol{P})$ and $(((q, p) \in \boldsymbol{P}^2) \Rightarrow ((\alpha p + (1-\alpha)q) \in \boldsymbol{P}$, $\alpha \in [0,1]))$ are realized, then the utility function *u*(.) is defined with precision up to an affine transformation $(u_1(.) \approx u_2(.)) \Leftrightarrow (u_1(.) = a u_2(.) + b,\ a>0)$ (in the case of utility function existence).

Now we are in interval scale and here the mathematical expectation is feasible. That is to say, this is a utility function [4, 7, 12]. The above properties related to Proposition 3 have also practical significance. This property is essential for the application of the utility theory, since it allows a decomposition of the multiattribute utility functions into simple functions [5]. Starting from the gambling approach for the definitions and the presentation of the expert's preferences we use the following sets motivated by Proposition 3:

$\boldsymbol{A}_{u^*}=\{(\alpha,x,y,z)/(\alpha u^*(x)+(1-\alpha)u^*(y))>u^*(z)\}$, $\boldsymbol{B}_{u^*}=\{(\alpha,x,y,z)/(\alpha u^*(x)+(1-\alpha)u^*(y))> u^*(z)\}$.

The notation $u^*(.)$ is the DM's empirical utility assessment. The approach we are using for the evaluation of the utility functions in its essence is the recognition of these sets. Through stochastic recurrent algorithms we approximate functions recognizing the above two sets [12]. Starting from the properties of the preference relation ($\succ$) and indifference relation ($\approx$) and from the weak points of the "lottery approach" we propose the next stochastic approximation procedure for evaluation of the utility function. In correspondence with the Proposition 3 it is assumed that ($\boldsymbol{X}\subseteq\boldsymbol{P}$), $((q,p)\in\boldsymbol{P}^2\Rightarrow(\alpha q+(1-\alpha)p)\in\boldsymbol{P}$, for $\forall\alpha\in[0,1])$ and that the utility function $u(.)$ exists. The following proposition is in the foundation of the used stochastic approximation approach:

Proposition4. We denote $\boldsymbol{A}_u=\{(\alpha,x,y,z)/(\alpha u(x)+(1-\alpha)u(y))>u(z)\}$. If $\boldsymbol{A}_{u1}=\boldsymbol{A}_{u2}$ and $u_1(.)$ and $u_2(.)$ are continuous functions than is true ($u_1(.)=au_2(.)+b$, $a>0$) [12]**.**

The approximation of the utility function is constructed by pattern recognition of the set $\boldsymbol{A}_u$ [12]. The proposed assessment process is machine-learning based on the DM's preferences. The machine learning is a probabilistic pattern recognition because ($\boldsymbol{A}_{u^*}\cap\boldsymbol{B}_{u^*}\neq\varnothing$) and the utility evaluation is a stochastic approximation with noise (uncertainty) elimination**.** Key element in this solution is the proposition 4. The following presents the evaluation procedure:

The DM compares the "lottery" $\langle x,y,\alpha\rangle$ with the simple alternative z, $z\in Z$ ("better-$\succ$", $f(x,y,z,\alpha)=1$", "worse-$\prec$", $f(x,y,z,\alpha)=(-1)$" or "can't answer or equivalent- $\sim$", $f(x,y,z,\alpha)=0$", $f(.)$ denotes the qualitative DM's answer). This determine a learning point ((x,y,z,α), $f(x,y,z,\alpha)$). The following recurrent stochastic algorithm constructs the utility polynomial approximation $u(x)=\sum_i c_i\Phi_i(x)$:

$$c_i^{n+1}=c_i^{n}+\gamma_n\left[f(t^{n+1})-\overline{(c^n,\Psi(t^{n+1}))}\right]\Psi_i(t^{n+1}),$$

$$\sum_n\gamma_n=+\infty,\ \sum_n\gamma_n^2<+\infty,\ \forall n,\gamma_n>0.$$

In the formula are used the following notations (based on $\boldsymbol{A}_{\mathbf{u}}$): $\boldsymbol{t}=(x,y,z,\alpha)$, $\boldsymbol{\psi}_i(t)=\boldsymbol{\psi}_i(x,y,z,\alpha)=\alpha\Phi_i(x)+(1-\alpha)\Phi_i(y)-\Phi_i(z)$, where $(\boldsymbol{\Phi}_i(x))$ is a family of polynomials. The line above the scalar product $\bar{v}=\overline{(c^n,\Psi(t))}$ means: ($\bar{v}=1$), if ($v>1$), ($\bar{v}=-1$) if ($v<-1$) and ($\bar{v}=v$) if ($-1<v<1$). The coefficients c_i^n take part in the polynomial presentation

$g^n(x) = \sum_{i=1}^{n} c_i^n \Phi_i(x)$ and $(c^n, \Psi(t)) = \alpha g^n(x) + (1-\alpha) g^n(y) - g^n(z) = G^n(x, y, z, \alpha)$ is a scalar product. The learning points are set with a pseudo random sequence. The mathematical procedure describes the following assessment process: The expert relates intuitively the "learning point" (x,y,z,α)) to the set $\boldsymbol{A}_{u^*}$ with probability $D_1(x,y,z,\alpha)$ or to the set $\boldsymbol{B}_{u^*}$ with probability $D_2(x,y,z,\alpha)$. The probabilities $D_1(x,y,z,\alpha)$ and $D_2(x,y,z,\alpha)$ are mathematical expectation of f(.) over $\boldsymbol{A}_{u^*}$ and $\boldsymbol{B}_{u^*}$ respectively, ($D_1(x,y,z,\alpha)$=M($f/x,y,z,\alpha$)) if (M($f/x,y,z,\alpha$)>0), ($D_2(x,y,z,\alpha)$=(-)M($f/ x,y,z,\alpha$)) if (M($f/x,y,z,\alpha$)<0). Let D'(x,y,z,α) is the random value: D'(x,y,z,α)=$D_1(x,y,z,\alpha)$ if (M($f/x,y,z,\alpha$)>0); D'(x,y,z,α)=(-$D_2(x,y,z,\alpha)$) if (M($f/x,y,z,\alpha$)<0); D'(x,y,z,α)=0 if (M($f/x,y,z,\alpha$)=0). We approximate D'(x,y,z,α) by a function of the type : G(x,y,z,α)=(αg(x)+(1-α)g(y)-g(z)), where $g(x) = \sum_i c_i \Phi_i(x)$. The coefficients c_i^n take part in the approximation of G(x,y,z,α): $G^n(x, y, z, \alpha) = (c^n, \Psi(t)) = \alpha g^n(x) + (1-\alpha) g^n(y) - g^n(z)$, $g^n(x) = \sum_{i=1}^{N} c_i^n \Phi_i(x)$. The function $G^n(x,y,z,\alpha)$ is positive over $\boldsymbol{A}_{u^*}$ and negative over $\boldsymbol{B}_{u^*}$ depending on the degree of approximation of D'(x,y,z,α). The function $g^n(x)$ is the approximation of the utility function u(.). It is used the following decomposition: $f(t^{n+1}) = \left[D'(t^{n+1}) + \xi^{n+1} \right]$.

The proposed procedure and its modifications are machine learning [12]. The computer is taught to have the same preferences as the DM. The DM is comparatively fast in learning to operate with the procedure, a session with 128 questions (learning points) takes approximately 45 minutes and requires only qualitative answers "yes", "no" or "equivalent ".

2.2 Optimal Portfolio Utility Allocation

This problem is discussed in different scientific sources and has practical significance [2, 6, 11]. But in all of them the problem of the choice (or of the construction) of the utility function is out of discussion. In our exposition we will propose a more complex utility function for description of DM's portfolio allocation. This utility (objective) function is constructed (approximated) by the stochastic procedure described in the previous subsection and is in agreement with the DM's preferences.

We use a classical dynamic model for description of a financial market. Consider a non-risky asset S^0 and risky one S. Following the sources [11] the Black-Scholes stochastic differential equation is given by:

$$dS_t^0 = S_t^0 r dt \quad \text{and} \quad dS_t = S_t \mu dt + \sigma dW_t .$$

Here r, μ and σ are constants (r=0.03, μ=0.05 and σ=0.3) and W is a one dimensional Brownian motion [11]. By X_t we denote the state space vector of the controlled dynamic process. The investment policy is defined by a progressively adapted process $\pi=\{\pi_t, t\in[0,T]\}$ where π_t represents (defines) the amount ($X_t \pi_t$, $\pi_t \in[0,1]$) invested in the risky asset at time t. The remaining wealth ($X_t - \pi_t X_t$) at the same moment t is invested in the risky asset. The time period T is 50 weeks. The liquidation value of a self-financing strategy satisfies [11]:

$$dX_t^{\pi} = \pi_t X_t^{\pi} \frac{dS_t}{S_t} dt + (X_t^{\pi} - \pi_t X_t^{\pi}) \frac{dS_t^0}{S_t^0} = (rX_t^{\pi} + (\mu - r)\pi_t X_t^{\pi}) + \sigma \pi_t X_t^{\pi} dW_t .$$

It is obvious that in the chosen parameters for the investment policy is true:

$$\mathrm{E} \int_0^T (\pi_t X_t^{\pi})^2 dt < \infty .$$

Here E denote mathematical expectation defined in the initial filtered probability space (Ω, F, F, P) with canonical filtration F=$\{F_t, t\geq 0\}$ of the Brownian motion defined over the probability space (Ω, F, P). More precisely E denotes mathematical expectation over the probability space (Ω, F, P). That is why the control process is well defined following theorem 2.1 of chapter 2 in [11]. The objective of the investor (DM) is to choose the control (the amount π_t invested in the risky process) so as to maximize the expected utility of his terminal wealth at moment T, i.e:

$$V(t,x) := \sup_{\pi\in[0,1]} \mathrm{E}[U(X_T^{t,x,\pi})],$$

Where $X^{t,x,\pi}$ is the solution of the controlled stochastic differential equation with initial condition (initial wealth) x at time t. For the liquidation value is supposed that if the state space vector is zero in a moment t then it remains zero to the end T ($X_T^{t,x,\pi} = 0$).

The optimal control could be determined step by step from the Hamilton-Jacobi-Bellman partial differential equation following the dynamical programming principle [1, 2, 10, 11]:

$$\frac{\partial w}{\partial t}(t,x) + \sup_{\pi\in[0,1]} [(rx + (\mu - r)\pi x)\frac{\partial w}{\partial t}(t,x) + \frac{1}{2}\sigma^2\pi^2 x^2 \frac{\partial^2 w}{\partial^2 t}(t,x)] = 0.$$

Following the presentation in the scientific source [6] and passing through generalized solution of the Black-Scholes stochastic differential equation we suppose that the solution of

the Hamilton-Jacobi-Bellman partial differential equation has the form $w(t, x)=U(x)h(t)$ where $U(x)$ is the DM's utility function.

Now is time to describe more precisely the utility function. We suppose that the total initial DM's wealth is 40000 BGN. But the DM does not wish to invest the whole initial sum in the risky process. We suppose that the DM invest in the risky process at moment $t\in[0,T]$ in accordance with his utility function if the wealth is between 0 and 40000 BGN. Over this sum (BGN) we choose the utility function of the form $U(x)=(U(40000)/40000^{\gamma})(40000+(x-40000))^{\gamma}$ where $x\in[40000, \infty)$ and ($\gamma=0.3$) according to [11]. Possible utility function is shown in figure 1 and 2 only for numeric demonstration of the approach.

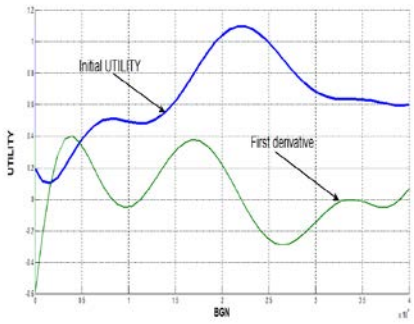


Figure1 Utility function between 0 and 40000 BGN

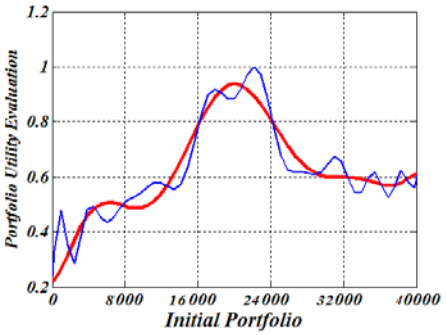


Figure 2 Utility construction

The determination of optimal control is easy because we know the first and the second utility function derivative analytically and the extremal point is looked for in the closed interval [0, 1] in a finite set of points. The optimal control $\pi(t_i)$ depends on $X_{t_i}^{0,x,\pi}$ and is the maximizer of

$$\sup_{\pi\in[0,1]}[(rx+(\mu-r)\pi x)\frac{\partial w}{\partial t}(t,x)+\frac{1}{2}\sigma^2\pi^2x^2\frac{\partial^2 w}{\partial^2 t}(t,x)] \ .$$

If the optimal control value belongs to the open interval (0, 1) then it has the following form

$$\pi(t_i)=\frac{-(\mu-r)U(X_{t_i}^{0,x,\pi})}{X_{t_i}^{0,x,\pi}U''(X_{t_i}^{0,x,\pi})\sigma^2}.$$

The investment policy (X_{ti} $\pi(t_i)$, $\pi(t_i)\in[0,1]$) is defined by the expressions

$\frac{-(\mu-r)U(X_{t_i}^{0,x,\pi})}{U''(X_{t_i}^{0,x,\pi})\sigma^2}$ if $\pi(t_i)\in(0,1)$ or 0 or $X_{t_i}^{0,x,\pi}$ in the other two cases.

The Belman function could not be determine analytically except in some special cases. In the following two figures (respectively 3 and 4) we show the solution with optimal control or without control.

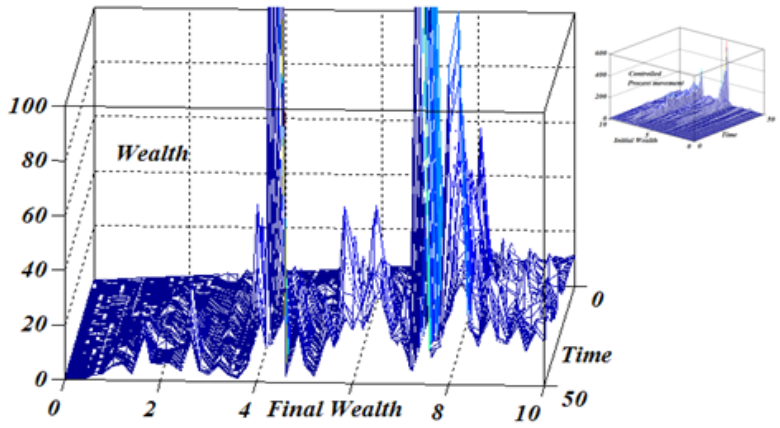


Figure 3 Liquidation value stochastic process with control

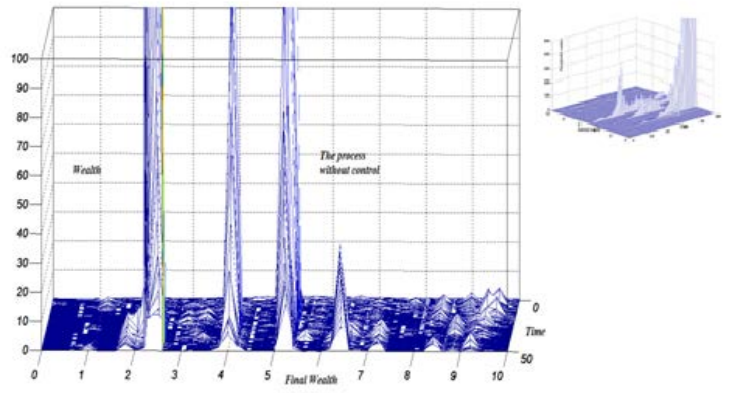


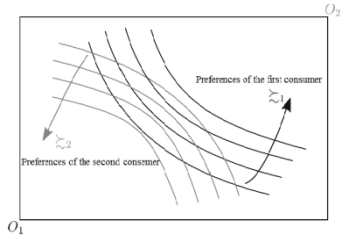
Figure 4 Liquidation value stochastic process without control

In figure 3 is seen the effect of optimal control strategy of the DM very well. This strategy is in accordance with his preferences presented analytically by the utility function.

2.3 Preferences and equilibrium in Edgeworth box model

The main purpose in this subsection is to be demonstrated the approach within determination of the equilibrium points in the competitive trading modeled by the Edgworth box [3]. A model for description the competitive trade is the Edgeworth Box. It merges the indifference map between the parties in the trade by inverting one of the agents diagram as is shown in figure 5. The demand functions or the utility functions which represent consumers'

preferences are convex and continuous and are shown in figure 5. Given two consumers $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$, two goods, and no production, all non-wasteful allocations can be drawn in the box shown in figure 6.



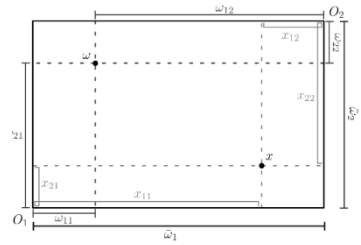
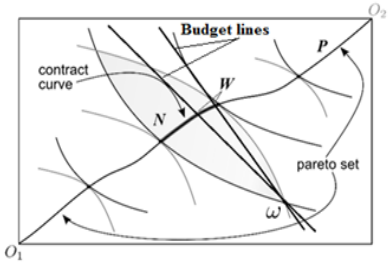


Figure 5. Consumer's demand's utility curves Figure 6. Edgeworth Box

Every point in the box represents a complete allocation of the two goods to the two consumers. Each of the two individuals maximizes his utility according to his preferences [3]. The demand utility functions (figure 5) which represent consumers' preferences are convex and continuous, because in accordance with the theory the preferences in are continuous, monotone and convex [3]. The two consumers are each endowed (born with) a certain quantity of goods. They have locally non-satiated preferences and initial endowments: ($\boldsymbol{w}_1$, $\boldsymbol{w}_2$) = ((w_{11}, w_{21}), (w_{12}, w_{22})). In the box the vector $w = (\overline{w}_1, \overline{w}_2)$ is the total quantities of the two goods: $\overline{w}_1 = w_{11} + w_{12}$, $\overline{w}_2 = w_{21} + w_{22}$. An allocation $\boldsymbol{x}$=($\boldsymbol{x}_1$, $\boldsymbol{x}_2$) = ((x_{11}, x_{21}), (x_{12}, x_{22})) represents the amounts of each good that are allocated to each consumer. A no wasteful allocation $\boldsymbol{x}$=($\boldsymbol{x}_1$, $\boldsymbol{x}_2$) is one for which is fulfilled: $\overline{w}_1 = x_{11} + x_{12}$, $\overline{w}_2 = x_{21} + x_{22}$. In terms of aggregate amounts of the two agents, the total amounts needs to be equal to the total endowment of the two goods. The consumers take prices of the two goods $\boldsymbol{p}$ = (p_1, p_2) as given and maximize their utilities. The budget (income) set $B_i(\boldsymbol{p})$ of each consumer is given by: $B_i(\boldsymbol{p}) = \{\mathbf{x_i} \in \mathbf{R}^2_+ \ / \ \boldsymbol{p}\mathbf{x_i} \le \boldsymbol{p}\boldsymbol{w_i}\}, (\mathbf{i} = 1,2)$, where ($\boldsymbol{p}\boldsymbol{x}_i$) and ($\boldsymbol{p}\boldsymbol{w}_i$) mean scalar products. For every level of prices, consumers will face a different budget set. The locus of preferred allocations for every level of prices is the consumer's offer curve.

An allocation is said to be Pareto efficient, or Pareto optimal, if there is no other feasible allocation in the Edgeworth economy for which both are at least as well off and one is strictly better off. The locus of points that are Pareto optimal given preferences and endowments is the Pareto set, noted as $\boldsymbol{P}$ in figure 7. The part of the Pareto set in which both consumers do at least as well as their initial endowments is the Contract curve shown in figure 7 and noted as $\boldsymbol{N}$ (kernel of market game).



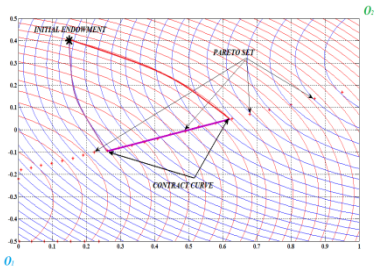


Figure7. Pareto and Walrasian set

Figure 8. DM"s contract curve

We are interested in the equilibrium point(s) of the process of exchange where is fulfilled the Walrasian equilibrium [3]. Walrasian equilibrium is a price vector $\boldsymbol{p}$ and an allocation $\boldsymbol{x}$ such that, for every consumer the prices (i.e. the terms of trade) are such that what one consumer (group of consumers) wants to buy is exactly equal to what the other consumer (group of consumers) wants to sell. In other words, consumers' demands are compatible with each other. We note the locus of points that are in Walrasian equilibrium as $\boldsymbol{W}$ (two points in figure 7). In still other words, the quantity each consumer wants to buy at the given market prices is equal to what is available on the market. The following inclusion is true in the Edgworth economy $\boldsymbol{P \supset N \supset W}$. In that sense a contract curve in the Edgeworth Box shows an exchange market in equilibrium and this is a particular representation of the Walrasian equilibrium theorem. The consumer's preferences are evaluated as value functions. In figure 8 are shown the indifference curves, the Pareto set $\boldsymbol{P}$ and the contract curve $\boldsymbol{N}$.

The indifference curves in figure 8 are determined based on values functions evaluated by direct comparisons of couples of allocations $\boldsymbol{x}=(\boldsymbol{x}_1, \boldsymbol{x}_2) = ((x_{11}, x_{21}), (x_{12}, x_{22}))$. This could be is made through the discussed in the paper approach and algorithms for exact value function evaluation ($\mathbf{A}_{u^*} \cap \mathbf{B}_{u^*} = \varnothing$). After that is made quadratic approximation of the constructed value function. The divergence from the theoretical convex requirements is due to the finite number of learning points and to the uncertainty in the expressed consumer's preferences. In the experiment for determination of the set $\mathbf{A}_{u^*}$ and $\mathbf{B}_{u^*}$we used a finite number of preferences expressed for couples of allocations ($\boldsymbol{x}=(\boldsymbol{x}_1, \boldsymbol{x}_2)$, $\boldsymbol{y}=(\boldsymbol{y}_1, \boldsymbol{y}_2)$):

$$\mathbf{A}_{u^*}=\{(x,y)\in \mathbf{R}^{2m} / (u^*(x))>u^*(y)\},\ \mathbf{B}_{u^*}=\{(x, y)\in \mathbf{R}^{2m} / (u^*(x))<u^*(y)\}.$$

In that manner we can state and solve the market-clearing equilibrium in principle and we can determine the contract curve and the Walrasian set in the Edgeworth box. The set of the Walrasian equilibriums $\boldsymbol{W}$ and the appropriate prices $\boldsymbol{p} = (p_1, p_2)$ are calculated based on the determined demand utility (value) functions and this is a meaningful prognosis of the

market equilibrium. In that way can be forecast the competitive market equilibrium allocations $\boldsymbol{x}=(\boldsymbol{x}_1, \boldsymbol{x}_2) = ((x_{11}, x_{21}), (x_{12}, x_{22}))$ and the appropriate prices $\boldsymbol{p} = (p_1, p_2)$. The contract curves are specified on the individual consumers' preferences and show that there are possibilities to be made mutually advantageous trades. This means that one could unilaterally negotiate a better arrangement for everyone.

3 Discussion

We wont to underline that the DM's utility could be constructed as a multiattribute utility function. In these conditions the numerical algorithm for solving the Hamilton-Jacobi-Bellman partial differential equation remains the same as is shown in subsection 2.2.

The determination of the equilibrium point (or the contract curve) in the Edgworth box in agreement with the DM,s preferences could made by value function also [3, 5]. The construction of DM's value and/or functions could be made by the algorithms described in [12].

4 Conclusions

In the paper is demonstrated a system engineering value driven approach within the problem of determination of the optimal portfolio allocation modeled with dynamic as Black-Scholes stochastic differential equation and the problem of determination of the equilibrium points in a competitive trading modeled by the Edgworth box. The Black-Scholes optimal portfolio solution and the contract curves (the trading equilibrium) in the Edgworth box are specified on the individual consumers' preferences. The mathematical formulations presented here could serve as basis of tools development. These value evaluation leads to the development of preferences-based decision support in machine learning environments and iterative control design in complex problems.

References

[1] Collopy, P. Hollingsworth, Value-driven design. *AIAA Paper* 2009-7099, *American Institute of Aeronautics and Astronautics*, Reston, VA, 2009.

[2] Crandall M.G., Ishii H., and Lions Pierre-Louis, (1992), User's Guide To Viscosity Solutions of Second Order Partial Differential Equations, *American Mathematical Society*, Volume 27, Number 1, July 1992, Pages 1-67.

[3] Ekeland, I., *Elements d'économie mathématique*, Hermann, 1979, Russian translation: Mir, 1983.

[4] Fishburn, P., *Utility theory for decision-making*, New York, Wiley, 1970.

[5] Keeney, R., H. Raiffa, *Decision with multiple objectives: Preferences and value trade-offs*, Cambridge & New York: Cambridge University Press, 1993.

[6] Mania M. & R. Tevzadze, (2008), Backward Stochastic Partial Differential Equations Related to Utility Maximization and Hedging, *Journal of Mathematical Sciences*, Vol. 153, No. 3.

[7] Pfanzagl, J., *Theory of Measurement*, Physical-Verlag, Wurzburg-Wien, 1971.

[8] Raiffa, H., *Decision Analysis*, New York: Addison-Wesley Reading Mass, 1968.

[9] Shmeidler, D., Subjective probability and expected utility without additivity, *Econometrica, 57*(3), 571-587, 1989.

[10] Smears Iain, (2010), Hamilton-Jacobi-Bellman Equations Analysis and Numerical Analysis. Report, Durham University, UK from http://maths.dur.ac.uk/Ug/projects/highlights/PR4/Smears_HJB_report.pdf

[11] Touzi N., Tourin Ag., (2012), *Optimal Stochastic Control, Stochastic Target Problems, and Backward SDEs*, Fields Institute Monographs, Vol. 29, Springer- Business & Economics.

[12] Pavlov Y. and R. Andreev, *Decision control, management, and support in adaptive and complex systems: Quantitative models*, Hershey, PA: IGI Global, 2013.

Big Data Approaches to Modeling the Labor Market

Anton Gerunov
Sofia University "St. Kliment Ohridski", Faculty of Economics and Business Administration
125 Tsarigradsko Shosse Blvd., 1115 Sofia, Bulgaria
gerunov@uni-sofia.bg

Abstract: The research paper leverages a big dataset from the field of social sciences – the combined World Values Survey 1981-2014 data – to investigate what determines an individual's employment status. We propose an approach to model this by first reducing data dimensionality at a small informational loss and then fitting a Random Forest algorithm. Variable importance is then investigated to glean insight into what determines employment status. Employment is explained through traditional demographic and work attitude variables but unemployment is not, meaning that the latter is likely driven by other factors. The main contribution of this paper is to outline a new approach for doing big data-driven research in labor economics and apply it to a dataset that was not previously investigated in its entirety.

Keywords: Labor market, Unemployment, Big data, WVS

5 Introduction

Traditionally econometric modeling has perused relatively small datasets to answer questions of substantive economic interest. A typical approach would be to formulate a scientific hypothesis, collect a limited number of theoretically-informed variables and subject them to statistical testing using largely linear models under the assumption of normality of the underlying data distribution [1]. This methodological framework has provided many fruitful insights and deepened our understanding of the underlying economic processes.

However, it suffers from a number of potential pitfalls –small samples raise questions about bias, estimation precision, and generalizability. Further, the number of observations imposes constraints on the maximum feasible number of independent model variables so that a researcher often has to make a judgment call on what to include. The availability of large datasets containing hundreds of millions or more data points ("big

data") can now help overcome those limitations and provide an additional perspective on economics research [2].

This paper focuses on modeling the labor market and proposes a way in which a well-known machine-learning algorithm can be applied to glean novel conclusions from a large-scale dataset. We first review traditional theories and methods for modeling the labor market and outline how big data approaches can supplement and enrich the existing paradigms. Then we fit an ensemble Random Forest model and outline the model qualities and main results. The paper concludes with directions for further research and possible applications of this very new approach in economics.

6 Labor Market Theories and Approaches

Labor market theories can be broadly subdivided into two main groups: microeconomic theories, and macroeconomic theories. Micro-level approaches emphasize the supply of labor as resulting from optimizing decisions by households and the demand for labor as resulting from optimizing decisions by firms. At the resulting equilibrium employment the firms pay for their workers a wage equal to their marginal productivity. A newer strand of theories – the "search and matching" theories – have focused much more on the process by which workers are matched to certain positions, given their useful economic characteristics like productivity, and employers' needs [3].

Macroeconomic approaches have tended to emphasize the connection between unemployment and the level of economic activity, thus intimately connecting labor market developments with economic growth and recessions [3]. The logic behind this is lucid – greater output is usually produced by an increase in inputs, and labor is one of the most important ones. Those two angles to understanding employment are of crucial importance – they emphasize firm-level needs, labor market processes, and macroeconomic structural needs. Some authors also point at the importance of psychological characteristics, values, and perceptions at the individual level for a given adult's employment prospects [4], [5]. The availability of large-scale data on individual characteristics allows us to see what values and attitudes determine employment in addition to a worker's productivity and job vacancies.

7 Data and Methods

To add the individual level dimension to modeling the labor market situation, we will explore sociological data coming from one of the largest social science primary data collection initiatives – the World Values Survey [6], and will outline a possible approach to gleaning insight from such large data for the purposes of economics research.

3.1 World Values Survey Dataset

The World Values Survey is an ongoing project since 1981 whereby respondents from almost 100 countries, covering 90% of world population, are questioned regarding their values, attitudes, beliefs, life conditions, demographic characteristics and evaluations. Currently about 330,000 respondents have been interviewed over the Survey's six waves and an additional seventh wave is presently under way. Separate variables from different waves have been used extensively by economists, psychologists and other social scientists to study in depth questions about political participation, economic development, culture and psychological issues. This wealth of data has never been analyzed in its entirety as the Survey Committee made publicly available the beta version of integrated and compatible data throughout all waves only very recently. This dataset contains 330,354 observations of 1377 variables for a total of nearly 455 million data points. While more common in machine learning contexts, such volumes of data are rarely studied in the social sciences. This paper will present a feasible approach to analyzing it in view of possible computational resource constraints. We will look in particular into the variable Employment.status, which codes whether the respondent is employed full-time (coded 1), part-time (2), self-employed (3), retired (4), housewife (5), student (6), unemployed (7) or in other position / not asked (codes 8 and negative). The WVS data will allow us to see what individual-level characteristics are important for classifying individuals in either of those positions, thus providing a useful exploratory analysis which can spur additional research at the boundary between economics, psychology and sociology.

Fig. 1: WVS Dependent Variable Distribution

3.2 Data Processing and Dimension Reduction

A dataset of such dimensionality – with a lot of variables per observation (so-called "fat data") – provides a clear computational challenge, which calls for optimization of its size. As with many big data sets, the World Values Survey is also a sparse dataset – some of the observations have either missing values or very low variance. Those are unlikely to be useful as classification and regression algorithms leverage variable variances. We remove them by using the function nearZeroVar as implemented in the caret packages of the R programming language, thus decreasing the number of variables to 485. A key insight from social science research is that a lot of the collected variables exhibit high multi-collinearity due to the complex feedback loops in social and economic systems. This means that some correlates can be dropped at the expense of very small information loss, so we check all pairs of variables with correlation of above r=0.9 and remove the one with the largest mean absolute correlation. Those two simple steps lead to a dramatic reduction of the number of variables per observation – these now count 233, or a total of 77 million data points. Such a dataset can be usefully analyzed with relatively fewer computational difficulties. For the analysis presented here we use the R Studio IDE on a server of 24 GB RAM and an eight-core Intel i7-4770 processor at 3.5 GHz, and will report calculation times for both sequential and parallel computation.

There are many other possible approaches for dimension reduction that economists may fruitfully apply in their big data modeling. For example, experimentation with Principal Component Analysis (or Singular Value Decomposition) proved suboptimal for the WVS

data but may be useful in other contexts. Large volumes of information also make it possible to select only a subsample from the data thus decreasing computation time.

3.3 Random Forest Ensemble

After pre-processing we can model the data by using a scalable machine-learning algorithm. Classification and Regression Trees (CART) are familiar to economists and are certainly one of the most useful and easier to interpret tools in the big data toolbox. There are also many standard and useful implementations but trees suffer from a number of problems, most notably high susceptibility to noise in data. Ensembling a large number of trees ameliorates this and effectively decreases the prediction or classification variance. Those ideas lead to the combination of trees into forests, so-called Random Forests, which can leverage a large number of trees for prediction purposes [7], [8]. Research in the machine learning literature has revealed that Random Forests show very good performance on a large number of problems under a wide range of specifications, making them a widely applicable algorithm. For this reason we will pursue modeling with them but naturally there are many other viable approaches [2].

The Random Forest algorithm proceeds as follows [7]: for ***b=1*** to ***B***, it selects a bootstrapped sample from the training data ***Z***. For our research, the size of this sample equals the whole dataset. Then, ***m*** variables are selected at random from the ***p*** predictors, so that ***m = sqrt(p)***. Out of those 15 variables the best variable/split-point is picked and the node is split into two daughter nodes, thus growing the tree T_b. We have limited the number of terminal nodes to a maximum of 10,000 for computational convenience. After all trees are grown (with ***B = 504***), those are combined in an ensembled Random Forest $\{T_b\}_1^B$. The prediction for a new point **x** in case of a regression Random Forest is then:

$$f_{rf}{}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

In case of classification, the trees in the ensemble generate a classification $C_b(x)$ and then "vote", and the majority vote assigns the class, or:

$$C_{rf}{}^{B}(x) = majority.vote\{C_b(x)\}_1^B.$$

We fit a classification random forest to the WVS data to investigate what variables are most important for classifying a respondent as Employed, Non-employed, Retired, Part-time Employed, Self-employed, Student, Housewife, etc. The model was sequentially calculated for 72.9 minutes. Since a Random Forest is particularly suitable for parallel

computation, we also split the 504 trees into 8 processes of 63 trees. The latter model was computed for 36.6 minutes, or an improvement of about 50%. Big data analytics of such scale reaps the benefits from parallelization and the algorithm selection needs to be performed with this possibility in mind. Reported results are for classification Random Forest, which is the more theoretically sound to apply. If one experiments with a regression Random Forest, the computation time increases to more than 20 hours sequentially and 8.3 hours in parallel.

8 Results and Discussion

The calculated model provides very good results, with an estimated out-of-bag (OOB) error rate of 39.29%, meaning correct classification of 60.71% of cases under scrutiny. Looking at the most important classification variables, we observe few surprises. Age, sex, marital status, ethnic group, education, and region top the list. Work attitudes which have captured the imagination of organizational theorists such as preference towards specific types of tasks (manual or cognitive) and preference for workplace autonomy are also present. The variables we observe are standard for the labor market literature and emphasize the importance of individual-level demographics well above attitudes in the job allocation process. In that sense the current exploratory study confirms results from previous research on the labor market.

One exception is the importance of God in one's life, which is slightly negatively correlated with better employment prospects. Atheists seem to be more likely to be employed full-time that the devout. A possible explanation could be their more pragmatic and less metaphysical focus. Such results naturally need to be interpreted with care and further investigated.

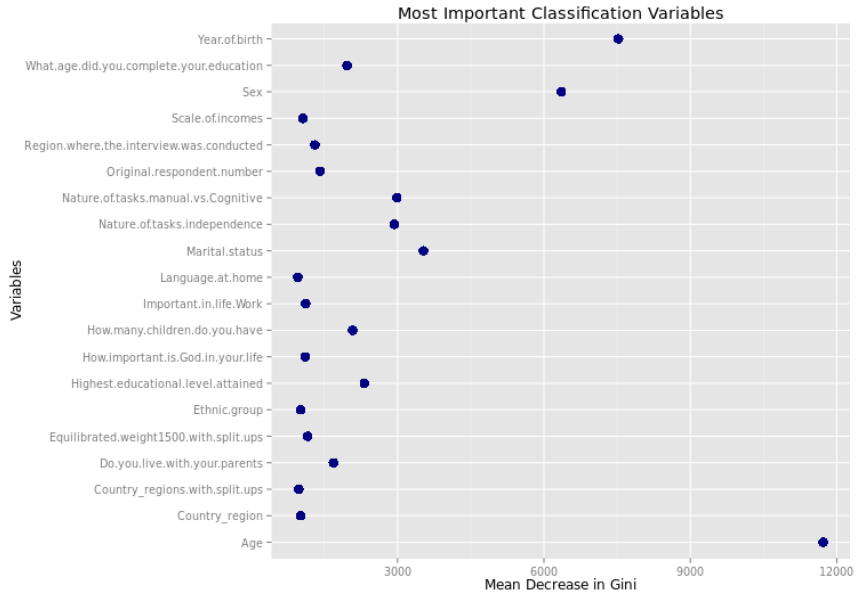


Fig. 2: Random Forest Classification Variables with Highest Mean Decrease in Gini

The variable with greatest predictive power by far is Age, underlining the strong life cycle effects on individual employment prospects. Full-time and part-time employed individuals (codes 1 and 2) tend to be of the same age, whereas self-employed (code 3) are slightly older. Retired (code 4) and students (code 6) are also easy to recognize. The unemployed (code 7) tend to be on average younger than those in employment. This piece of statistics is likely driven by the fact that unemployment has larger prevalence among young adults than in the general population.

The abridged model confusion matrix is presented in Table 1. Individuals in full-time employment are largely correctly classified, showing that there are some very distinct characteristic within this group. Other groups in the labor force such as self-employed, or part-time employed are largely classified as Employed, meaning that these are only little different in attitudes and perceptions from those actually having a full-time job.

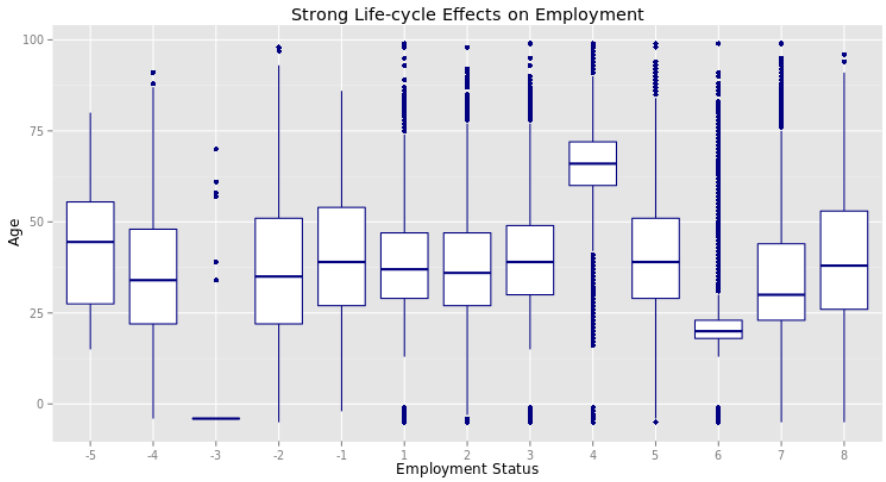


Fig. 3: Median Age of Individuals Across Different Employment Statuses

Unemployed individuals are difficult to classify correctly. One possible interpretation of this result is that there is hardly a stable set of psychological attitudes, social beliefs or evaluations that clearly distinguish the unemployed from the other respondents. These results underlie the fact that reasons for unemployment will need to be found in individual labor productivity, motivation, and overall job market conditions, rather than in a set of psychological beliefs and attitudes.

Table 1: Truncated Confusion Matrix for Random Forest Classification

Class / True	*Full-time*	*Part-time*	*Self-employed*	*Retired*	*Housewife*	*Student*	*Unemployed*	*Other*	*Error Rate*
Full-time	102604	30	1152	2791	3352	1899	260	58	8.6%
Part-time	19807	187	630	1381	2128	1352	114	84	99.3%
Self-employed	23079	9	6070	2041	2618	634	208	129	82.6%
Retired	6493	4	145	29224	1646	57	202	28	22.8%
Housewife	12032	24	294	3040	32893	603	313	63	33.4%
Student	5987	0	21	48	576	17745	482	2	28.7%
Unemployed	15874	2	526	1418	3450	3698	3826	57	86.7%
Other	3067	12	165	689	583	518	236	571	90.2%

The results so far underline an interesting conclusion – we are very well aware what makes a person employed – active age, good education, being in the right ethnic group and the right region of the country, and possessing pro-work attitudes. What remains elusive is what makes a person unemployed – even those with favorable characteristics might end up without a job, and we seem to be unable to statistically distinguish between the former and the latter using individual demographics and attitudes. In that sense the current paper opens interesting venues for labor market research.

Firstly, employment and unemployment do not seem to be the flipsides of the same coin, as is commonly assumed in labor economics, but rather two distinct conditions that need to be studied separately. Secondly, demographics, psychological attributes, and social perceptions seem unable to explain unemployment and other explanatory factors need to be investigated further. An obvious determinant of unemployment is individual labor productivity which probably plays a role. Another viable contender is chance. If there is a structural labor market need for downsizing the labor force, some individuals may lose their jobs purely by chance, irrespective of their objective qualities. While this interpretation substitutes randomness for causality, it might be worth exploring further.

Thirdly, such results can be uniquely gleaned only through leveraging a combination between big data and advanced machine learning algorithms. Under the standard econometric inference testing approach one could utilize a version of the Generalized Linear Model to interpret regression coefficients and their significance levels. This will only show that some regressors reach statistical significance, and are therefore important for predicting the dependent variable. We will not be able to see the subtle differences and discern that employment and unemployment are two very different conditions that may need to be studied within distinct theoretical and analytical frameworks.

9 Concluding Remarks

The current exploratory study leverages a new and previously unutilized dataset – the complete integrated comparable World Values Survey data spanning 1981-2014 – to investigate if individual level employment can be explained by a combination of demographics, psychological attributes, and social attitudes. Using a Random Forest model we classified respondents, with over 60% out-of-bag correct classification. Employed individuals were largely correctly classified, but the unemployed ones were more challenging. A possible reading of this result is that unemployment is hardly defined by

traditional individual level attributes (age, education, region, work attitudes) but could be attributed to either individual labor productivity or structural labor market characteristics and randomness of outcomes. Such results can serve to refocus the research agenda in labor economics and steer it towards a better understanding of the determinants of individual employment status.

References

[1] Greene, W. (2011). *Econometric Analysis, 7th Edition*. US: Prentice Hall.

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The Elements of Statistical Learning*. NY: Sprigner.

[3] Romer, D. (2012). *Advanced Macroeconomics*. US: McGraw-Hill.

[4] Kalil, A., Schweingruber, H. & Seefeldt, K. (2001).Correlates of Employment Among Welfare Recipients: Do Psychological Characteristics and Attitudes Matter? *American Journal of Community Psychology*, Volume 29, Issue 5 , 701-723.

[5] Kessler, R. C., Turner, J. B., & House, J. S. (1987). Intervening processes in the relationship between unemployment and health. *Psychological Medicine*, *17*, 949–961.

[6] World Values Survey, Wave 1-6 1981-2014. (2014). World Values Survey Association (www.worldvaluessurvey.org).

[7] Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), 5-32.

[8] Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2-3, 18-22.

Performance analysis of a load-frequency power system model

Svetoslav Savov, Ivan Popchev
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Acad. G. Bonchev Bl. 2, 1113 Sofia, Bulgaria
savovsg@mail.bg

Abstract: This research works investigates the derivation of a fixed upper matrix bound for the solution of one class of parameter-dependent continuous algebraic Lyapunov equation (CALE). It is supposed that the nominal coefficient matrix is subjected to a real structured parametric uncertainty belonging to a convex set. The bound is used to analyze the robust stability and the performance behavior of a load-frequency control system for a single area power system model. By means of the bound one can easily estimate the distance from instability of the uncertain system and the associated with it linear quadratic performance index. The applicability of the obtained results is illustrated by an example.

Keywords: Lyapunov equation, solution bounds, uncertain systems, power systems.

1. Introduction

The problem of deriving bounds for the solution of the CALE attracts interest for more than half a century. This is due to both theoretical and practical reasons. In some cases, due to high order the direct solution of this equation is impossible, and in other ones, it is sufficient to have at disposal only some estimates for it. The main difficulty arises from the fact, that the available upper bounds are valid under some assumed restrictions imposed on the coefficient matrix. Due to this, valid solution bounds are possible only for some special subsets of negative stable (Hurwitz) coefficient matrices. All significant results in the area are summarized and discussed in [9].

Robustness of a linear system, subjected to structured real parametric uncertainty, belonging to some compact vector set (e.g., the unit simplex), has been recognised as a key issue in the analysis of control systems [1-6]. The main purpose of this research is to derive fixed upper matrix bound for the solution of one class of parameter-dependent CALEs. Such bounds help to analyze the uncertain system with respect to stability and a quadratic

performance index. A state space model with real data of a single area power system is used as test example.

The following notations will be used: $A > (\geq) 0$ indicates that *A* is a positive (semi-) definite matrix; $\alpha = (\alpha_i) \in \mathbf{R}^{\mathbf{N}}$ denotes a real *n* x 1 vector α with nonnegative entries α_i, $i = 1,..., N$ and $|\alpha|$ is the sum of its entries; $A^{1/2}$, A^{-1}, A^T are the square root (if *A* is positive semi-definite), the inverse (if *A* is nonsingular) and the transpose of a matrix *A*; $\lambda_m(A)$, $\lambda_M(A)$ denote the minimal and maximal eigenvalue of a matrix *A* with only real eigenvalues, respectively; the real part of the i-th eigenvalue of matrix *A* is $\mathrm{Re}\,\lambda_i(A)$; v^* is the conjugate transpose of a complex vector *v*. All matrices are *n* x *n*. The identity matrix is denoted *I*. Define also the set of *n* x *n* uncertain matrix polynomials:

$$\boldsymbol{P} \equiv \{A(\alpha), \quad \alpha = (\alpha_i) \in \mathbf{R}^{\mathbf{N}}: \quad A(\alpha) = A + \sum_{i=1}^{N} \alpha_i A_i, \quad 0 \leq |\alpha| \leq 1\}$$

where A, A_i, $i = 1,..., N$, are some fixed matrices.

2. Preliminaries

Consider the state space model of a linear continuous-time uncertain system:

$$\dot{x} = A(\alpha)x, \quad x(0) = x_0, \quad A(\alpha) \in \boldsymbol{P} \tag{1}$$

and the associated with it parameter-dependent CALE

$$A^T(\alpha)P(\alpha) + P(\alpha)A(\alpha) = -Q, \quad Q > 0 \tag{2}$$

From Lyapunov's stability theorem it follows that if $A(\alpha)$ is a Hurwitz (negative stable) matrix for all admissible vectors α, i.e.

$$s(\alpha) = -\max \mathrm{Re}\,\lambda_i[A(\alpha)] > 0, \quad i = 1,.., n \tag{3}$$

then $P(\alpha)$ is the unique positive definite solution of equation (2) for any given positive definite matrix *Q*. In this case, the performance of the system can be evaluated by the index:

$$J(\alpha, x_0) = \int_0^\infty x^T(\alpha)Qx(\alpha) = x_0^T P(\alpha)x_0 \tag{4}$$

It is desired to determine parameter independent bounds for the:

(a) positive definite solution $P(\alpha)$ in (2)

(b) distance from instability $s(\alpha)$ in (3)

(c) system performance index $J(\alpha, x_0)$ in (4)

Before that, the following simple results will be presented.

Lemma 1. A symmetric uncertain polynomial $X(\alpha) \in \boldsymbol{P}$ is positive definite if and only if it is positive definite at the $N + 1$ vertices, i.e.

$$X > 0, \quad X + X_i > 0, \quad i = 1,..., N \tag{5}$$

In this case, for the positive scalar

$$\mu = \max\{\lambda_M(QX^{-1}), \quad \lambda_M[Q(X + X_i)^{-1}], \quad i = 1,..., N\} \tag{6}$$

one has

$$\mu X(\alpha) \geq Q, \quad \forall \alpha \tag{7}$$

Proof. Suppose that $X(\alpha)$ is a positive definite polynomial for all α. Then, the matrix inequalities in (5) must hold, by necessity, which proves the necessity part. Now, let the set of matrix inequalities (5) holds. Since the sum $|\alpha|$ of the entries of vector α belongs to the interval [0, 1], there always exists some nonnegative scalar α_0, such that $\alpha_0 + |\alpha| = 1$. This results in the $N + 1$ matrix inequalities:

$$\alpha_0 X \geq 0, \quad \alpha_i(X + X_i) \geq 0, \quad i = 1,..., N$$

with at least one of them being strict. By summing up the left and right-hand sides one gets

$$X + \sum_{i=1}^{N} \alpha_i X_i = X(\alpha) > 0, \quad \forall \alpha$$

This proves the sufficiency part and completes the proof of the first statement.

Let $X(\alpha)$ be a positive definite polynomial for all α and consider the scalar defined in (6). Obviously, its choice guarantees that:

$$Y = \mu X - Q \geq 0, \quad Y + Y_i = \mu X - Q + \mu X_i \geq 0, \quad i = 1,..., N$$

Application of the same arguments used to prove the first statement, one gets that the inequality (7).

Lemma 2. Let $A(\alpha)$ be a Hurwitz matrix for all α. If there exists a fixed symmetric matrix P_U, such that

$$A^T(\alpha)P_U + P_U A(\alpha) \le -Q, \quad \forall \alpha$$

(8)

then P_U is an upper parameter independent matrix bound for the solution $P(\alpha)$ in (2), i.e. $P(\alpha) \le P_U, \forall \alpha.$

Proof. If the above inequality holds, having in mind (2), one has:

$$A^T(\alpha)[P_U - P(\alpha)] + [P_U - P(\alpha)]A(\alpha) \le 0, \quad \forall \alpha$$

This is possible only if $P_U - P(\alpha) \ge 0, \forall \alpha$, in accordance with Lyapunov's stability Theorem.

Corollary 1. If the assumptions of Lemma 2 hold, then having in mind the scalars in (3) and (4), one gets the following parameter independent bounds for the distance from instability and the performance index of the uncertain system (1):

$$s(\alpha) \ge s = \frac{1}{2}\lambda_n(QP_U^{-1}), \quad \forall \alpha \tag{9}$$

$$J(\alpha, x_0) \le J(x_0) = x_0^T P_U x_0, \quad \forall \alpha, \forall x_0 \tag{10}$$

Proof. Let γ denotes an eigenvector of $A(\alpha)$ corresponding to the eigenvalue λ with the largest real part for all uncertain vectors α i.e. $A(\alpha)\gamma = \lambda\gamma$. Consider the matrix inequality (8) and the associated with it scalar inequality:

$$\begin{aligned} \gamma^* Q\gamma &\le -\gamma^*[A^T(\alpha)P_U + P_U A(\alpha)]\gamma \\ &= -(\lambda^*\gamma^* P_U\gamma + \lambda\gamma^* P_U\gamma) \\ &= -2\,\mathrm{Re}(\lambda)\gamma^* P_U\gamma \\ &= 2s(\alpha)\gamma^* P_U\gamma \end{aligned}$$

Note that P_U must be a positive definite matrix by necessity. Finally, denoting $\varphi = P_U^{1/2}\gamma$ results in the inequality

$$\lambda_n(QP_U^{-1}) = s \le \frac{\varphi^* P_U^{-1/2} Q P_U^{-1/2}\varphi}{2\varphi^*\varphi} \le s(\alpha)$$

This proves the bound in (9). The upper bound (10) is obvious.

3. The Power System Model

The purpose of operating load-frequency control is to keep the frequency changes during the load sharing in some desired limits. The main change parameters of a power system are the rotor angle, the change in frequency and the active power flow between the connection lines. The given below linear model is taken from [7] and is sufficient to express the dynamic behavior of the system around the working point [8]

$$\Delta\dot{P}_v = -\frac{1}{\tau_g}\Delta P_v - \frac{1}{R\tau_g}\Delta f$$

$$\Delta\dot{P}_m = \frac{1}{\tau_T}\Delta P_v - \frac{1}{\tau_T}\Delta P_m$$

$$\Delta\dot{f} = \frac{1}{2H}\Delta P_m - \frac{D}{2H}\Delta f - \frac{1}{2H}\Delta P_L$$

where the respective parameters have the following physical meanings:

$\Delta P_v, \Delta P_m, \Delta f$ denote the change in turbine valve position power, turbine mechanic exit power and frequency, respectively; τ_g, τ_T are the speed regulator time constant and the turbine time constant, respectively; H and D denote the generator inertia constant and the power system constant. The control input is ΔP_L and denotes the load change. Using the notations

$$x = \begin{pmatrix} \Delta P_v & \Delta P_m & \Delta f \end{pmatrix}^T, \quad u = \Delta P_L$$

the system is put in the standard state space description of an open-loop system

$$\dot{x} = A_0 x + bu, \quad A_0 = \begin{bmatrix} -\frac{1}{\tau_g} & 0 & -\frac{1}{R\tau_g} \\ \frac{1}{\tau_T} & -\frac{1}{\tau_T} & 0 \\ 0 & \frac{1}{2H} & -\frac{D}{2H} \end{bmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{2H} \end{pmatrix}$$

The following values for the parameters are used:

$$\tau_g = 0.2\ s,\ \tau_T = 0.5\ s,\ R = 0.05\ \text{pu},\ H = 5\ s,\ D = 0.08$$

The state and control matrices are computed as:

$$A_0 = \begin{bmatrix} -5 & 0 & -100 \\ 2 & -2 & 0 \\ 0 & 0.1 & -0.008 \end{bmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \\ -0.1 \end{pmatrix}$$

Although the open-loop system is stable, a procedure of an optimal linear quadratic regulator synthesis is suggested via the solution of the algebraic Riccati equation:

$$-\tilde{Q} = A_0^T R + RA_0 - Rbb^T R$$

$$= (A_0 - bb^T R)^T R + R(A_0 - bb^T R) + Rbb^T R$$

$$= (A_0 - bK)^T R + R(A - bK) + Rbb^T R$$

$$= A^T R + RA + Rbb^T R$$

The close loop system with a state feedback control law $u = -Kx,\ K = b^T R$, becomes

$$\dot{x} = (A_0 - bK)x = Ax \qquad (11)$$

where *R* denotes the Riccati equation solution. The close loop system state matrix *A* satisfies a Lyapunov-like equation

$$A^T R + RA = -Q, \quad Q = Rbb^T R + \tilde{Q} \qquad (12)$$

The Riccati equation has been solved for $\tilde{Q} = diag.(10,\ 7,\ 3)$. The gain matrix has been computed as

$$K = \begin{pmatrix} 1.0595 & -0.1099 & -45.271 \end{pmatrix}$$

Now, we want to investigate the robustness properties of the nominal system (11) by including the action of an additive structured polynomial perturbation. i.e.

$$\dot{x} = Ax + \sum_{i=1}^{N} \alpha_i A_i x = A(\alpha)x, \quad A(\alpha) \in \boldsymbol{P}$$

This puts the uncertain system in the form (1). A sufficient condition for its stability is due to Lyapunov's stability theorem, which in a case when a fixed Lyapunov function is required, is given by the matrix inequality:

$$0 > A^T(\alpha)R + RA(\alpha)$$

$$= A^T R + RA + \sum_{i=1}^{N} \alpha_i (A_i^T R + RA_i)$$

$$= X + \sum_{i=1}^{N} \alpha_i X_i$$

$$= X(\alpha)$$

Since $X(\alpha) \in \boldsymbol{P}$, having in mind (12), and in accordance with Lemma 1, this equality has a simple parameter independent solution:

$$-X = Q > 0, \quad -X - X_i = Q - A_i^T R - RA_i > 0, \quad i = 1,..., N \quad (13)$$

Let $N = 2$ and

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 20 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 30 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

All three matrices in (13) are positive definite, which guarantees negative definiteness of $X(\alpha)$, stability of the uncertain state matrix $A(\alpha)$ and of the system for all admissible vectors α.

Let the right-hand side in the parameter-dependent CALE (2) be the identity matrix. The scalar μ in (6) has been computed as follows:

$$\mu = \max(0.1429, 0.1602, 0.144) = 0.1602 = \lambda_M[(-X - X_1)^{-1}]$$

According to Lemma 1, one has

$$-I \geq \mu X(\alpha)$$

$$= A^T(\alpha)(\mu R) + (\mu R)A(\alpha)$$

$$= A^T(\alpha)P_U + P_U A(\alpha)$$

Due to Lemma 2, this means that

$$P_U = \begin{bmatrix} 0.1665 & 0.0609 & -1.6970 \\ 0.0609 & 0.2886 & 0.1760 \\ -1.6970 & 0.1760 & 72.5121 \end{bmatrix}$$

is a fixed upper matrix bound for the parameter-dependent solution of the CALE (2) for all α. From (9) one gets

$$s(\alpha) \geq s = \frac{1}{2}\lambda_n(P_U^{-1}) = 0.0069, \quad \forall \alpha$$

An upper estimate for the performance index in (10) can be easily computed for any given initial state vector.

References

[1] Bliman P. (2004) A convex approach to robust stability for linear systems with uncertain scalar parameters. *SIAM Journal of Control and Optimization*, 42, 6, 2016-2042.

[2] Bliman P. (2004) An existence result for polynomial solutions of parameter-dependent LMIs. *Systems and Control Letters*, 51, 3, 165-169.

[3] Chesi G., Garulli A., Tesi A., Vicino A. (2005) Polynomially parameter-dependent Lyapunov functions for robust stability of polytopic systems: an LMI approach. *IEEE Transactions on Automatic Control*, 50, 3, 365-379.

[4] Geromel J., Korogui R. (2006) Analysis and synthesis of robust control systems using linear parameter dependent Lyapunov functions. *IEEE Transactions on Automatic Control*, 51, 12, 1984-1988.

[5] Grman L., Rosinova D., Vesely V., Kozakova A. (2005) Robust stability conditions for polytopic systems. *Journal of System Science*, 36,15, 961-973.

[6] Oliveira R., Peres P. (2007) Parameter-dependent LMIs in robust analysis: characterization of homogeneous polynomially parameter-dependent solutions via LMI relaxations. *IEEE Transactions on Automatic Control*, 52, 7, 1334-1340.

[7] Kakilli A., Oguz Y., Calik H. (2009) The modelling of electric power systems on the state space and controlling of optimal LQR load-frequency. *Journal of electrical and electronics engineering* 9, 2, 977-982.

[8] Saadat H. (1999) Power system analysis. McGraw Hill, New York.

[9] Savov, S. (2014) Solution bounds for algebraic equations in control theory. Prof. M. Drinov Academic Publishing House, Sofia.

Primary information preprocessing system for LP, DP devices - project "Obstanovka"

Dichko Bachvarov*, Ani Boneva*, Bojan Kirov**, Yordanka Boneva*,
Georgi Stanev**, Nesim Baruh***
* Institute of Information and Communication Technologies - BAS, Acad. G. Bonchev Str., Bl. 2, Sofia, Bulgaria, dichko1952@abv.bg
** Institute of Space Research –BAS, Sofia, Bulgaria, bkirov@space.bas.bg
*** ELL "Danev,Bojilov" Ltd. - Sliven, Bulgaria

Abstract: The article presents Primary pre processing information system designed for using with Bulgarian devices LP and DP, working on ISS. There are described Bulgarian activities in the project "Obstanovka", the conversion process from telemetry to science data, LP and DP data structures, software solutions and system realisation. It is developed method for multi machine processing of big data areas. The article is illustrated by presentation of pre processed real science data of LP experiments.

Keywords: Space research, ISS, Obstanovka, LP, DP, science data, big data areas processing.

1. Introduction

The international space station (Fig.1 – International Space Station) is the most significant and the most expensive international space project so far. The project is collaboration between the Russian Academy of Science, NASA, the European Space Agency and Japan. Since April 2013 on the board of the Russian segment of the ISS (Fig, 2) the international experiment "Obstanovka" has begun. Its first phase includes the building of plasma-wave complex for measuring of wave and plasma parameters in the surrounding environment of the station. Bulgaria participates in the project by the creation of equipments for measuring the parameters of low-temperature plasma. In the experiment "Obstanovka" also other institutes and specialists from Russia, Ukraine, Poland, Hungary, Sweden and England participate. Since 17.04.2013 the Research complex has been put on the board of the ISS and after its activation on 27.04.2013 the scientific experiments have started.

2. Scientific research equipment in the project “Obstanovka”

The participation of Bulgaria in the project includes the development and conduction of scientific research with four devices that work in open space – 2 devices from type DP for measurement of the disturbance of the electromagnetic field in the space near the station and two devices LP (Fig.4) with the purpose to research the parameters of low-temperature plasma by using the method of Longmuir (Fig.2) [1]. These devices, together with other specialized devices are installed in two containers, fixed by masts located on different distances from the board of ISS (Fig.5). Each container represents separate automatic measurement complex that conducts various physical measurements under the same conditions and same time. The two containers are part of a local Ethernet network together with a specialized BSTM computer and a client computer on the board of the station. (Fig.6). BSTM supports a database of scientific measurements on portable disk.

3. Organization of the scientific data [2]

The data in BSTM is presented in primary files that include different types of records produced by different scientific devices together with internal terminal tags and meta- data. . The structure of these files is general – they include synchronizing fields, temporal information, meta- data fields and records with variable length that contains research data. The measurements of particular device records have different formats - the format is defined by the device’s developer. Besides, it is possible to exist a difference between experiments produced by one and the same device(different working modes can use different representation of the data). Things become more complicated because of the fact that particular experiment with particular device can produce a huge amount of data that make sense as a whole (i.e. separate parts of the data are not informative). Usually these records do not form a continuous flow – they are sequential for the specified device but between these records there are also records from other devices i.e. it is nessesary to isolate the information coming from a specified device and to assemble the whole structure of the data in a particular experiment.

3.1. Organization of scientific data in the LP

The LP device can accomplish 6 different scientific experiments each of them having different presentation of data as set of elements (structural units). The length of the structure that describes an experiment varies in number of elements (for the different experiments). To

ensure the integrity and accuracy of the information in the third element a 32-bit CRC of the data is included. In this element the type together with the given parameters of the experiment are specified. The first and second elements contain synchronization sequences. The fourth element containes the time of the experiment start, as measured from the beginning of the calendar year (in time units from 0.01 sec.). Next set of elements describes the results of the scientific experiment.

3.1 Organization of scientific data in the DP

The DP device can perform one scientific experiment having 3 parameters. Its data is grouped into elements similar to LP. 16 consecutive elements form a block. Depending on the value of the 3rd parameter, the structure of experiment length may be 1, 2, 4 or 8 blocks. Each block contains 32 bit CRC to control the integrity, as well as a time to start the block of measurements.

3.2 Visualization of scientific data

Each experiment allows graphical and tabular visualization of the data included in its structure. This can be done, provided that all the data included in the structure of the experiment is correctly read. In case of errors, the data is ignored and the experiment is considered invalid.

4. System for pre-processing of data from scientific experiments with LP and DP devices [3]

Obtaining and analyzing the results of experiments with the LP and DP devices require the design of specialized software package allowing a researcher access to the observed physical quantities. This package, called pre-processing system must implement the following functions:

- Access to primary files sent by telemetry channels in the ISS contact center. These files are stored in FTP server in ISR – Moscow. The package should have a built-in FTP client, through which the operator can control remotely the access to them(Fig.3.);
- Retrieving of information associated with a specific device from given primary (.dat) file and generate of the corresponding text file containing all logically consistent data records;
- Processing of text file, and forming the data included in the structure of the individual experiments;

- Recording this data (for each experiment separately) in the database for the given device, indexed by the time of the experiment;
- Other requirements for the pre-processing system include visualization of results. The package contains a set of software tools that allow the researcher monitoring of data from a separate experiment in graphic form (using built-in graphics library), in tabular form or as a file with export options in another package (Excel);
- There is also "Group" mode in which the data of the experiments included in the set by the operator time interval are provided in a format and are saved to a file suitable for export to an external package for further processing.
- The package is implemented as application under Windows XP / 7, named Obstanovka17, and after installation it provides the operator with a graphical interface (API)). On Fig. 6 the main screen of this application is shown, and Fig. 7 and Fig. 8 present the results of experiments with real LP and DP devices.

5. Parallel processing of incoming data area

The structure of Science Results Data Base supposes abilities for parallel processing of the incoming data areas.

Each “experiment_file” of measurements data of concrete instrument unit (LP1,LP2,DP1,DP2) is placed (into the DTBS folder) corresponding to the time of experiment starting. The file name includes the type of experiment processed with the unit.

There are two different base folders – one (LP1dtbs) for LP1 and one (LP2dtbs) for LP2. Each of them has tree hierarchy organization with 6 hierarchy levels (corresponded to year, month, day, hour, minute and seconds). By this way the file places into Database are sorted naturally. The lowest level (second) gives name of a folder, including corresponded “experiment_file” and empty “number_file“. The name of the last corresponds to the ordered number of the “experiment_ file”.

For example:

/LP1dtbs/14/10/01/10/01/00/(DE_2_2_1.txt ,1001.num),

here the text “14/10/01/10/01/00/” determinates the date and time of the experiment starting, DE_2_2_1.txt is “experiment_file” and 1001.num gives us the number of “experiment_file” into the sorted list of all files in LP1dtbs.

It is supporting of ordered file names list of included files into corresponded DTBS at current time. The adding of “experiment_file” into the “DTBS” forces creation of new entry into file structure or writing over existed one.

There are realized two function for searching of files into DTBS:

- By name of the “experiment_file” and the time of it creating to determinate it ordered number;
- By ordered number of the file to determine it file name and the time of it creation.

Each time when primary file processing is finished is resorting the file names of DTBS list and renaming of the “number_file” associated with the “experiment_file”.

Used data organization allows parallel processing of data-stream on two or more machines and monitoring at this time on other computer of scientific data.

It is possible next scenario:

- The group of machines includes pre-processing ones and special scientific computers;
- On all of them are installed program packages Obstanovka17;

After starting and initialization of it all machines automatically create own DTBS file structure (empty initially);

- Each of pre-processing machines is connected to FTP server hosted in ISR RASc. (Moscow) and executes the task of receiving of set of primary files, after then pre-process them und update its local DTBS. It is doing in parallel of all preprocessing computers. At end of preprocessing task, each local DTBS is converting into self-extracted archive.

The researcher has ability to get these compressed local DTBS and move them to his Research computer, where they are extracting automatically and adding to global DTBS. All actions connected to internal file processing and transport are supporting by program package Obstanovka17.

6. Conclusion

The data from different experiments with the Bulgarian LP and DP devices allow recording of physical events and monitoring of the parameters of low temperature plasma near the the board of ISS. The Primary processing system provides opportunities for monitoring and processing of the information obtained in the laboratory for the work period (4 years) and conducting systematic and thorough research of the results.

References

[1] KIROV B., Batchvarov D, Krasteva R, Boneva A., Nedkov R, Klimov S, Valery Grushin V. ,(Sopron, 24-29 August, 2009), ("LANGMUIR PROBES FOR THE INTERNATIONAL SPACE STATION",306-THU-P1700-03160 *AGA 11th Scientific Assembly* .

[2] Kirov B,*, Georgieva K , Batchvarov D , A. Boneva , Krasteva R, Stainov G , Klimov S, Dacheva T (2008) "Remote upgrading of a space-borne instrument" *Advances in Space Research* 42, 1180–1186

[3] Batchvarov D., B. Kirov, A. Boneva, R. Krasteva, S. Klimov, K. Georgieva, Software Package for Primary Processing of Telemetric Information, Tenth Jubilee National Conference wich International Participation Dedicated to the 70th Anniversary of Acad. Dimitar Mishev, *Contemporary Problems of Solar-Terrestrial Influences*, Proc. ISBN 954-91424-1-8, Editor: Acad. Dr. Stoycho Panchev, 6. Session "Space Instrumentation and Technologies"-SIT, STIL-BAS, Sofia, 20-21 November 2003, pp. 202-205.

Figure 1. ISS and "Obstanovka" project.

Figure 2. Bulgarian Space Instruments.

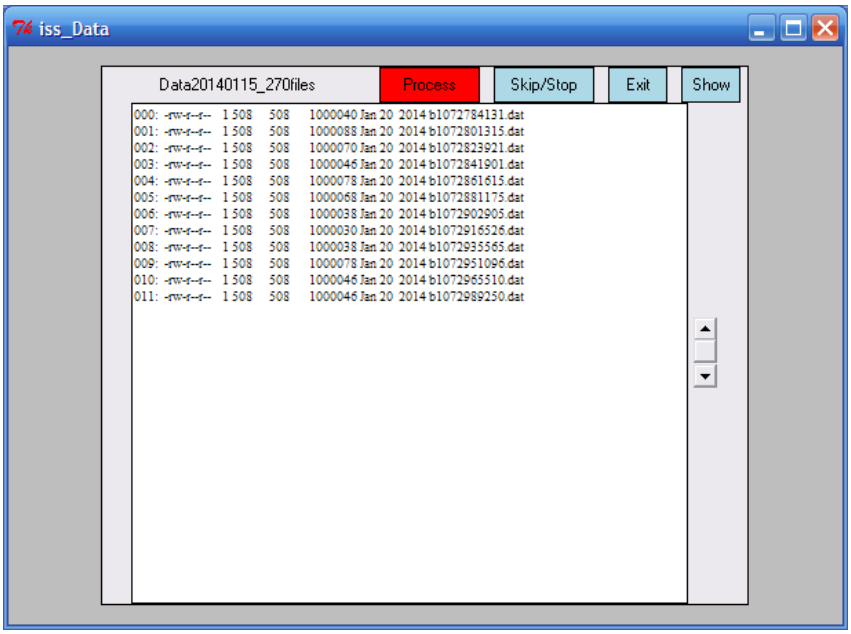


Fig.3. FTP Control Panel of OBSTANOVKA17.

12.9.2013 г.

Figure 4. Bulgarian Space instruments on the Board of ISS.

12.9.2013 г.

Figure 5. Network topology of “Obstanovka” Framework.

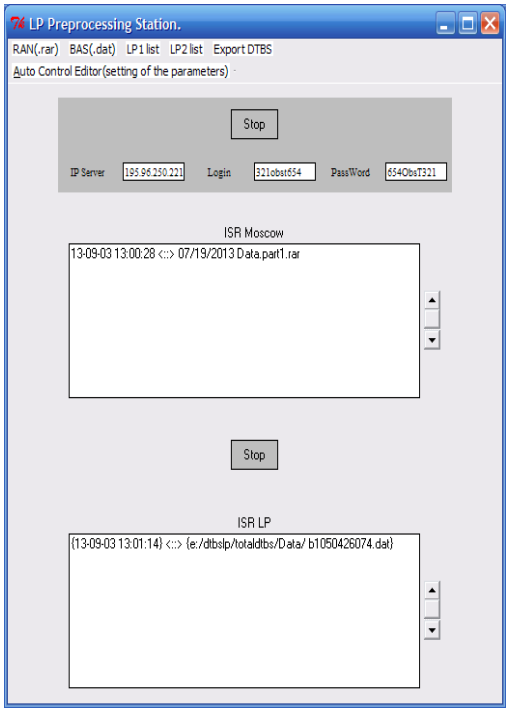


Figure 6. Main screen.

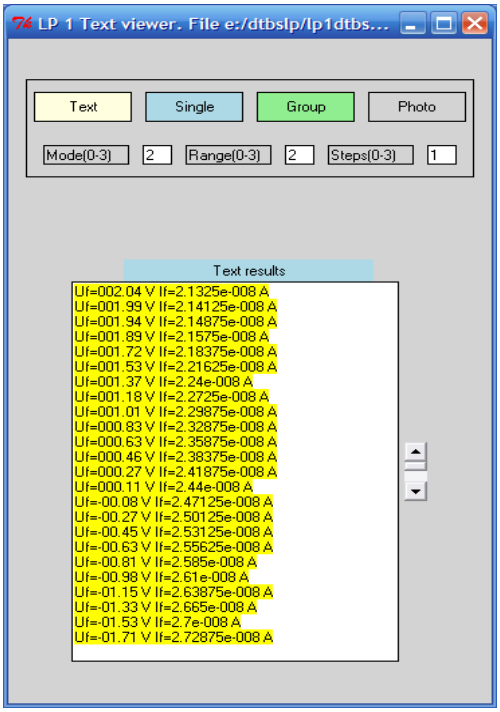


Figure 7. Tabular representation.

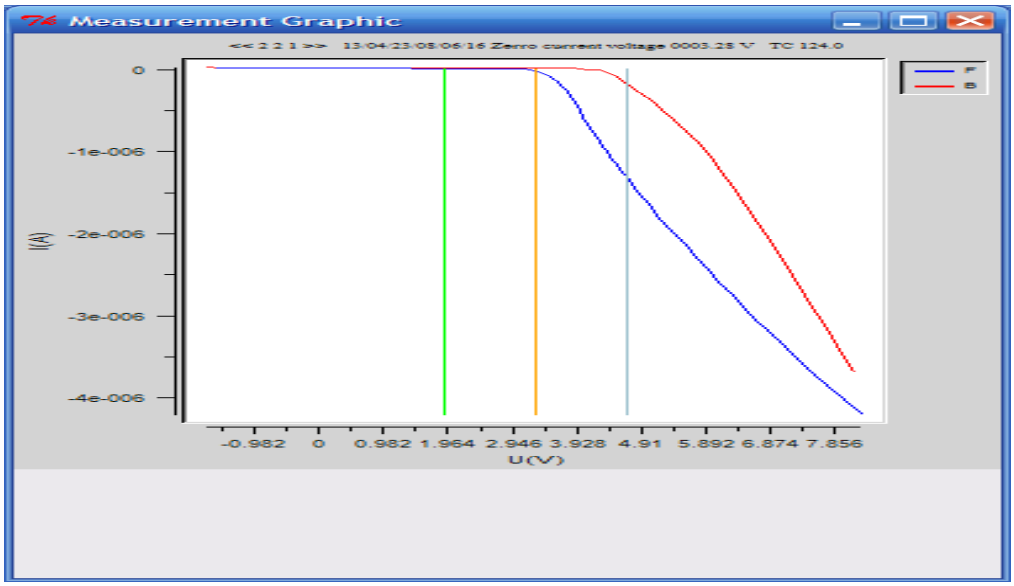


Figure 8. Graphical representation

Measurement Analysis that Defines Burner Operation of Hot Water Boilers

Milena Todorovic, Dragoljub Zivkovic, Marko Mancic, Pedja Milosavljevic, Dragan Pavlovic

Faculty of Mechanical Engineering, University of Nis

14 Aleksandra Medvedeva, Nis, Serbia

milenatod1@yahoo.com

Abstract: The paper presents an analysis of measured results of combustion process of hot water boilers that use fuel oil and natural gas, depending on the load of boiler. The paper focuses on parameters that should be controlled in flue gas and determination of their specific limits in order to have combustion process with the highest coefficient of utilization and also satisfying environmental requirements. The measurements were done with a digital instrument for measuring temperature, relative humidity, air flow, differential gas (reference instrument) with associated accessories and printer type TESTO 350M product Testo GmbH with the possibility of measurement of O2, CO2 , NO, NO2, temperature of the gases, ambient temperature, complete with accessories, printers, and probe, with the ability to store data and software. The measurement was carried out on hot water boilers manufactured by " Djuro Djaković" - Slavonski Brod with 5.37 MW and 16.5 MW capacity within the Heating plant of city of Nis.

Keywords: Combustion process, hot water boiler, measurement results, boiler start up, burner operation.

1. Introduction

Intensive research work has been dedicated to energy savings in recent years, which especially takes into account air pollution and lack of energy fuel. The research for this purpose have shown that efficiency of boilers, such as industrial or boilers in thermal power plant, which use fossil fuels, represent very important parameter. Any improvement of boiler efficiency leads to energy savings and economic viability of primary energy use [3, 6, 9, 10, 11].

In order to increase the efficiency of boiler, the combustion efficiency represents very important parameter. Research shows that boiler efficiency increases by 5% during complete combustion process compared to the process where we have incomplete combustion [1, 15].

It was done a lot of research about boiler systems so far that use different types of fuel. When the temperature of flue gases get lower on the outlet of the boiler, before chimney pipe, boiler efficiency increases, and also during the process of operation, heat losses and fuel consumption decreases.

Large number of industrial boilers that use natural gas for combustion were designed and built at time when fuel prices were relatively low, and often were derived for alter-native combustion of fuel oil and natural gas. Changes of fuel prices have caused producers of boilers to correspond to changes in structural details in order to reduce heat loss from outgoing flue gases during the natural gas combustion.

The paper focuses on the parameters that should be controlled in combustion products and identify some of their limits in order to have combustion process with the largest energy efficiency while environmental requirements will be satisfied. The paper suggest approaches that allow increasing of effectiveness of using natural gas and fuel oil in boiler plants, which represent the largest consumer of this kind of fuel. This primarily relates to lower temperature of combustion products at the outlet of the boiler during the operation of combustion unit with optimal excess air. At the end it is given an overview of the measurement results of composition and temperature of combustion products that is done using flue gas analyzer Testo 350M. The measurement was done on water boilers produced by "Djuro Djaković" - Slavonski Brod, with burner that use oil fuel and natural gas.

2. Parameters Defining Complete Combustion Process

With fuel combustion is released, with finite velocity, certain amount of hat that is transferred via combustion products and transformed into other forms of energy. As fossil fuels don't represent inexpensive source of energy, it is necessary keep this losses down to minimum, and to achieve the same energy effect with less fuel consumption. Given that in our country there is a widely developed net-work of consumers of natural gas and crude oil, their optimal operation is necessary, especially in terms of security, economy and ecology. Accordingly, optimization of combustion process in order to rationally fuel consumption refers to [2, 4, 8]:

- Controlled combustion (to obtain amount of heat required for the process);
- Burning of fuel with the highest level of efficiency;
- The least possible environmental pollution.

Combustion represents a chemical process of binding combustible constituents of fuel with oxygen from air, with heat deliverance. Depending on the amount of the oxygen

brought into process, combustion may be complete or incomplete. In general, when complete combustion occurs, combustion products are: CO2, H2O, SO2, NOx, N2 and O2. When incomplete combustion occurs, in addition to complete combustion products, there are also fuel components, which, if combustion process were complete, they would entirely give their amount of heat which they contain. Products of incomplete combustion are: CO, CmHn, H2, C. When burning of fuel oil occurs, due to presence of sulfur S in the fuel, one of additional combustion products is SO2.

Theoretically, combustion process will always be complete, if amount of oxygen, which is brought into the process is greater than or at least equal to the minimum of the required amount of oxygen for complete combustion.

Also, one of the influential parameters affecting the quality of combustion process is the burning rate. Burning rate must be equal to the velocity of propagation of the mixture in order to have steady flame and quality combustion. The maximum combustion rate occurs at stoichiometric conditions, while with increase of excess air or with deficit of air, burning rate decreases. The most important part of burner, which affects on the quality of the fuel-air mixture, is burner tube with nozzle and the mixing chamber.

Great impact on the efficiency has coefficient of excess air. It defines the amount and compositions of combustion products and amount of heat that they carry. When the temperature of the combustion products increases, the energy efficiency decreases (the heat losses are increasing), assuming that the content of CO2 and O2 in combustion products does not change. Also, reducing the coefficient of excess air (in the range of optimum combustion), at constant temperature of combustion products causes efficiency increasing. This temperature should be in the range of 160-220 °C, which is measured by standard methods on specified places behind the boiler.

Thus, it can be said that the ratio of excess air represents the main parameter that defines the quality of the combustion process. The lower the coefficient of excess air is, the higher the percentage of CO2 and the smaller proportion of O2 in the flue gasses are, the lower is the heat loss and thus higher energy utilization.

Besides the high efficiency, it must be satisfied another criterion that is a minimum of environmental pollution, and that the content of harmful substances CO and NOx in combustion products to be within acceptable limits. Be-tween these two requirements we must find a compromise. Excess air should be as lower as possible, but such that the content of CO and NOx in flue gases are within permissible concentration.

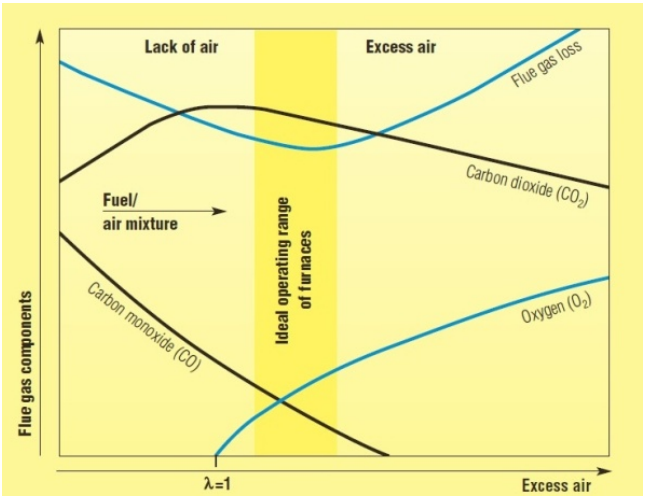


Figure 1. Diagram of optimal combustion [2]

Therefore it can be concluded from the above men-tioned that the main parameters that should be controlled in combustion products in order to achieve optimum combustion: content of oxygen, carbon dioxide, carbon mon-oxide, oxides of nitrogen and the temperature of the combustion products.

When we have incomplete combustion, which can occur due to lack of oxygen or poor mixing of combustible gases with air or hypothermia due to flammable gases, flue gases contain even more and unburned components, especially carbon monoxide CO and H2 as well as char. Due to high heating value of CO, only small content of CO in gases represents a significant loss of heat. Measurement of CO and H2 in the flue gases in the combustion chamber is therefore important for operating control.

3. Combustion Parameters Measurement in Dependence of Boiler Load

From the composition of flue gases it can be evaluated the quality of combustion. Therefore, in well-guided and operated combustion chambers composition of flue gases is continuously controlled by means of special measuring instruments. The most favorable ratio of excess air is the one in which occurs the lowest heat losses. The highest content of CO2 in flue gas is not favorable, because when it occurs often occur and carbon monoxide CO. Used measuring device is a digital instrument for measuring temperature, relative humidity, velocity of differential gas (reference instrument) with associated accessories and printer type Testo 350M produced by Testo GmbH with the possibility of measuring the content of O2, CO2, NO, NO2, temperature of flue gases, ambient temperature, complete with

accessories, printers and measuring probes, with the ability to archive data and suitable software [12].

The measurement was carried out on hot water boilers manufactured by "Djuro Djaković Slavonski Brod" with capacity of 5,37 MW and 16,5MW within the Heating plant in the city of Niš (Tables 1, 2).

Table 1 - Technical characteristics of the boilers [13, 14]

	Boiler 1	**Boiler2**
Manufacture:	"Djuro Djaković" - Slavonski brod	"Djuro Djaković" - Slavonski brod
Type:	Optimal 800	Optimal 2500
Maximum capacity of boiler:	5,37 MW	16,96 MW
Permitted maximum overpressure:	12,5 bar	16,2 bar
Operating pressure:	12,5 bar	15,7 bar
Temperature of hot water at inlet:	90°C	100°C
Temperature of hot water at outlet:	130°C	160°
Total heated surface:	136,5 m2	434,7 m2
Surface area of the flame:	8,5 m2	22,5 m2
Irradiated surface of the fire tube:	3 m2	25,4 m2
Surface of water-cooled front:	5,3 m2	19,2 m2
Surface of gas pipes of second pass:	61 m2	209,9 m2
Surface of gas pipes of third pass:	58.7 m2	157,7 m2
Amount of water in boiler:	10,845 m3	40 m3
Boiler efficiency:	87%	91%

Table 2 - Technical characteristics of the burners [13, 14]

	Boiler 1	**Boiler2**
Manufacture:	"SAACKE" - Germany	"SAACKE" - Germany
Type:	SKVJG 55-18	SKVG-A 82
Nominal capacity:	6,6 MW	17,3 MW
Fuel:	Fuel oil, gas	Fuel oil, gas

During the process of starting up the boiler as well as the process of shutting up, the composition values of flue gases were measured in dependence of the boiler load. The measured value of shares O2 and CO2, as well as the coefficient of excess air and combustion efficiency are presented for each boiler on figures 2, 3, 4 and 5.

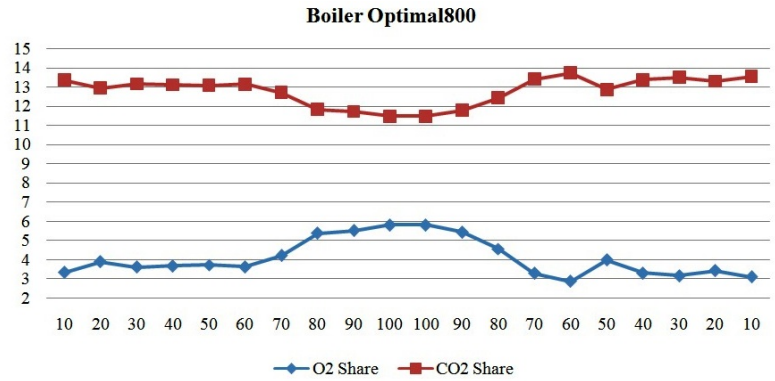


Figure 2. The Share of O2 and CO2 in flue gases depending of the boiler load for boiler Optimal 800

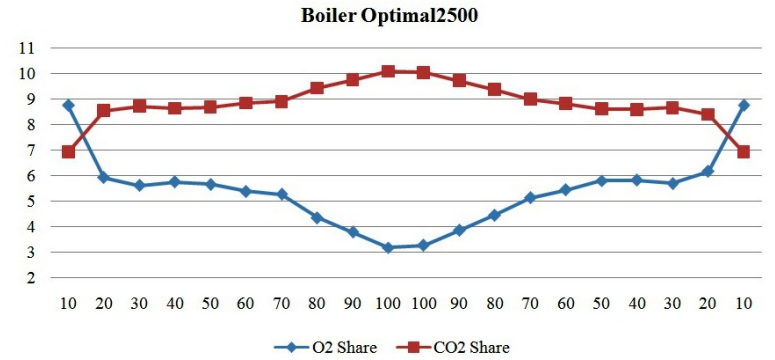


Figure 3. The Share of O2 and CO2 in flue gases depending of the boiler load for boiler Optimal 2500

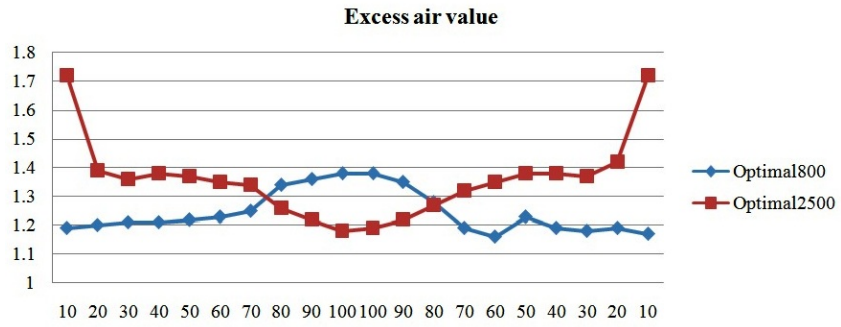


Figure 4. Excess air value λ depending of the boiler load for boilers Optimal 800 and Optimal 2500

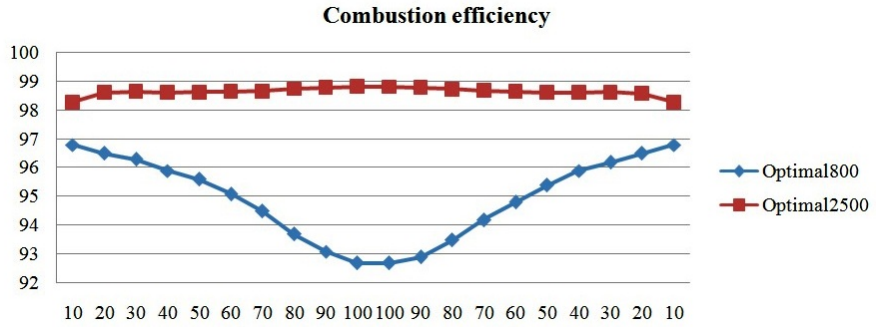


Figure 5. Combustion efficiency η depending of the boiler load for boilers Optimal 800 and Optimal 2500

Discussing results it can be said that regarding the boiler Djuro Djaković Optimal 800 (if we observe Fig.2) at higher loads there is incomplete combustion, because the share of O2 increases, while the share of CO2 decreases as we approach the load of 100%. Also it can be seen that the values of excess air coefficient λ that are higher at higher loads, so that at boiler load of 90-100% are 1.36-1.38, that immediately indicates that we have excess air that appears in the fuel-air mixture, as well as inefficient set the burner. Optimal combustion process takes place at boiler load of 50-70%. Also it was observed during the measurement at higher loads (80-90%) existence of CO, which indicates that we have incomplete combustion and the occurrence of carbon monoxide as very harmful gas.

Regarding the boiler Djuro Djaković Optimal 2500 (Fig.3), the situation is completely different. At higher loads of 80-100% of the boiler, the share of O2 and CO2 corresponds to the optimal mixture and closing to the values of the stoichiometric combustion. As can be observed also based on the value of the coefficient of excess air λ which at that load amounts 1.18-1.22. Excess air coefficient is the main indicator by which it can be evaluated the rationality of fuel combustion. It can be said that keeping the boiler load on the 80-90% corresponds terms of quality and complete combustion.

4. Conclusion

For effective use of fuel, natural gas or fuel oil, as well as heat that is produced during its combustion, it is necessary to control the combustion process by analyzing the combustion products. It is enough to determine the content of the flue gases (the share of CO2, O2 and CO).

The first indicator of incomplete combustion is the appearance of carbon monoxide in the flue gases, which is usually accompanied by darker color of the flue gas at the outlet of

the chimney. This is generally the consequence due to insufficient amount of air, or unsatisfactory mixing natural gas with air.

The temperature of the combustion products at the out-let of the boiler, for particular fuel, mainly depends on the type, size and age of the boiler and can reach the value of 250-300 °C and even higher for smaller boilers older structures, especially in the case of boilers, which were modified from solid or liquid fuel to natural gas. Temperature of the flue gases of boilers fired by natural gas are usually moving into the area to 150-180 °C, and the excess air 1.3-1.4 [7].

In the considered case, depending on the boiler load, the temperature of flue gases at the outlet of the boiler is at range from 90-160 °C regarding boilers fired with oil fuel and for the boilers fired with natural gas 100-165 °C.

When we have burning of fuel oil it must be maintained also slightly higher temperature of flue gases because the appearance of low-temperature corrosion caused by the presence of sulfur in the fuel oil. When we have natural gas combustion, the temperature of combustion products can be much lower (even below 100 °C) since natural gas does not contains sulfur or its components.

It may be noted, in this particular case, that the temperature of flue gases in hot water boiler fired by natural gas higher than temperatures of flue gases in hot water boiler fired by fuel oil. It is observed that it is possible to utilize this waste heat of the flue gases by setting appropriate economizer. This would be very successful use of the heat from outgoing flue gases which would irretrievably gone into atmosphere.

Regarding the occurrence of CO in flue gases, which would indicate to incomplete combustion, CO occurs in the smaller amount during the combustion of fuel oil, and these values are quite small 1 ppm and 2 ppm at 90% and 100% boiler load. Which also confirms the statement shown in the paper [5] that the carbon monoxide CO is fast intermediate which is formed at higher temperatures, where it is in significant quantities, and moved very quickly to CO2 in colder zones.

If however in the combustion products are present combustible gases (CO, H2 and CH4), which indicates to in-complete combustion, which usually happens when there is not enough professional exploitation of boilers, which has resulted in foiling of heat exchanges surfaces, resulting in reduced heat transfer and increases of flue gas temperature. Thus, increasing the heat losses with flue gases at outlet of the boiler, that is, decreases the coefficient of boiler efficiency.

Today exploitation of hot water boilers shows that there are a large number of boilers, especially smaller ones, without gas analyzer. Even there where the analyzers were built they are usually not used. Taking into account the economic effects that are achieved by monitoring the combustion process and affordability of flue gas analyzers which are available on the market, today every industrial boiler or boiler in thermal power plant should be equipped with exhaust gas analyzer. It is particularly important to organizations that perform the process of adjustment of the combustion process use combustion products analyzers.

If we have data of composition of the flue gases that are leaving the boiler as well as the parameters that define the optimal combustion, it can be determined the optimal ratio of fuel-air which is necessary to avoid unfavorable situation in the future, such as:

- That incomplete combustion will not be repeated;
- Boiler will be working for longer period of time on the most rational way without the intervention of the service staff.

References

[9] Ayhan B., Demirtas C. (2001) Investigation of Turbulators for Fire Tube Boilers Using Exergy Analysis. Turkish Journal of Engineering and Environmental Sciences, 25, 4, 249-258.

[10] Bogner M., Scekic G. (2011) Combustion chambers and burners. Eta, Belgrade.

[11] Brkic Lj., Zivanovic T., Tucakovic D. (2010) Steam boilers. University of Belgrade, Faculty of Mechanical Engineering, Belgrade.

[12] Djuric V. (1969) Steam boilers. Gradjevinska knjiga, Belgrade.

[13] Dramlic D., Miocinovic D. (1997) Semiconductor Sensors in Combustion Control of Gaseous Fuels. Proceeding of the GAS, Budva.

[14] Leizerovich A. (2008) Steam turbines for modern fossil-fuel power plants. The Fairmont Press, Inc., Lilburg.

[15] Perunovic P., Pesenjanski I. (1997) Increasing the Effectiveness of Natural Gas Utilization in Boiler Plants. Proceedings of the GAS, Budva.

[16] Recknagel H., Sprenger E. (1982) Heating and air conditioning. Gradjevinska knjiga, Belgrade.

[17] Sijacki Zeravcic V., Bakic G., Djukic M., Andjelic B. (2007) Analysis of Test Results of Hot-Water Boiler as a Basis for Its Integrity Assessment. *Structural Integrity and life*, 7, 2, 133-140.

[18] Sekeljic P., Bakic G. (2007) Optimization of Maintenance Measures of Piping System of The 60MW Boiler In Order To Raise Their Availability. *Energy-Economy-Environment*, 3-4, 45-49.

[19] Taler J., Michalczyk K. (2006) Thermal and Structural Stress Analysis of Boiler Drum and Central Pipe Connection in Transient Conditions. *AGH University journals - Mechanics*, 25, 1, 41-46.

[20] Technical documentation and user's guide for Testo 350M/XL.

[21] Technical documentation of boiler "Djuro Djaković"- Slavonski brod, Oprimal 800.

[22] Technical documentation of boiler "Djuro Djaković"- Slavonski brod, Oprimal 2500.

[23] Van Wylen G.J., Sonntag R.E. (1985) Fundamentals of classical thermodynamics. John Wiley and Sons, New York.

Big Data – an Essential Requisite of Future Education

Valentina Terzieva, Petia Kademova-Katzarova, Katia Todorova
Institute of Information and Communication Technologies – BAS
Acad. G. Bonchev Str., Bl. 2, Sofia, Bulgaria
valia@isdip.bas.bg, petia@isdip.bas.bg, katia@isdip.bas.bg

Abstract: The paper reveals the potential of the Big Data concept implemented in education. There are various sources for accumulating digital data in educational institutions at different stages. It is shown how the power of innovative tools for data collection, management and analysis are used for identification of good practice or problems in educational process. Considering this findings, some suggestions for providing individualized teaching are debated. An example scenario for generation, accumulation and interaction of Big Data in educational context is described. The possible benefits and drawbacks of Big Data employment in education for society and individuals are discussed.
Keywords: Big Data, Education, Learner's profile

1 Introduction

Big Data is a term that emerges from the growth of the information in the digital age in all organizations and it has potential to reveal relations when analyzing the data. Big Data refers to the huge amount of information from various sources – digital and conventional. It is too large, complex, and dynamic so that conventional tools do not manage to process it. The term emerged recently when advancements in ICT allow making sense of this data, providing great benefits to the data-driven economics, education and society as a whole. IBM suggests three different dimensions of Big Data, which include volume, velocity, and variety; later three more are added. The main characteristics of Big Data are described as follows [1]:

- **Volume** (huge quantity of data – generated, stored, analysed, managed)
- **Variety (**various sources, categories, formats, sizes)
- **Velocity** (the speed of generation and processing of data)
- **Variability** (frequent actualisation and alterations of data can lead to inconsistency)
- **Veracity** (truthfulness of the source data)
- **Complexity** (elaborative data processing and management)

This paper is focused on the ways that Big Data can be used to improve student success and educational process as a whole. The above mentioned main characteristics give us reasons for grounded implementation in educational context, which is described in details in the next part. Part three give some ideas for implementation of Big Data techniques in education sector. Part four describes how Big Data impact adaptive and personalised learning. The paper concludes with identifying the benefits of application of Big Data methods in education.

2 Big Data in education

2.4 Characteristics of Big Data in education

Nowadays, there is a great variety of educational environments – not only the traditional classroom, but various e- learning systems, also other education-related sources such as: learning resource repositories, concept maps, social networks, forums, educational games etc. provide different educational data, thus enabling different issues to be managed using Big Data techniques. This exponential growth of educational data from diverse sources needs adequate analysis handling and control in order to meet the objectives and specific goals of education [2]. For the successful development of the society well educated individuals are of vital importance. Therefore a constant stream of information on the effectiveness of the educational institutions is required [3]. For this purpose institutions have to keep track and maintain a records about all components of the educational process (students' profiles and performance; learning resources; curriculum). In order to gather, analyze, and manage this huge amount of information new processing methods have to be discovered and applied. It turns out that those used in data mining and Big Data can be effectively applied to education-related data.

In education Big Data are "big" both in size and significance because they allows the analysis and interpretation of wide range of information – demographic, personal (background, skills, learning style, psychological type, academic performance, IQ level, etc.), environmental factors and so on. In other words – anything that can be measured can be used as potential data for management of educational process. In our opinion, the use of Big Data can lead to quality improvement in education but the application is still at the very beginning in Bulgaria. In educational aspect these characteristics can be interpret as follows:

1. Volume – Big Data can give information about all learners taking some learning course and for numerous educational resources over time. Big Data should be scalable, so they

might bring together the data from many sources (schools, universities, organizations) to present a global view on education process.

2. Velocity – Big Data enables teachers and stakeholders almost in real time to collect, process and access data for management of the educational process and decision making. The students can access stored data for their own learning performance, approach learning resources, make assessments and receive in real time exam results.

3. Variety – Big Data encompass a wide variety of information for students: knowledge background, cognitive abilities, learning path, and portfolio. It allows spotting various correlations in the educational process.

4. Variability – Big Data allows keeping track of alterations and variation in learning processes over time.

5. Veracity – Big Data provides authenticity and reliability of the education data.

6. Complexity – Big Data allows management and process of very complex, large volume data from multiple sources. They are linked, connected and analysed in order to derive implicit information.

Considering this Big Data characteristics, we suggest to design educational strategies and to evaluate their impact on all participants in education, in order to enforce a grounded approach, and personalize the learning process.

2.5 Big Data analytics in education

Educational data can be structured and unstructured and usually are generated from many heterogeneous sources mainly offline (from traditional classroom and academic education – concern information about curriculum, learning courses, learners' attendance and performance, interaction among learners themselves and educators, etc.) or online sources (from interactive learning environments – distance, web-based and computer-supported education systems, social networks, thematic online forums, groups of interest, blogs, etc.) [4] The stakeholders analyse deeply this large amount of educational data in order to make well-grounded decisions, to build development strategies and predictions for the entire educational system.

The approaches applied for data processing are based on well known data mining methods. Data mining employs various tools and techniques for analysing and revealing relationships and patterns that are explicitly and implicitly included in the data set. It helps organizations in discovering essential information that assists them in decision-making processes. The applied methods are based on machine learning, artificial intelligence, computer science, and

statistics [5]. The most used techniques are clustering, classification, association rule, sequential pattern analysis, dependency modelling, etc. [6]. They are employed to analyze both quality and quantity aspects large data sets in order to uncover and explore patterns and trends in ongoing processes.

At first, processing educational data came into view from the analysis of logs when students interact with computer-based resources. Romero and Ventura in [7] categorize methods for management of this data as follows:

- ✓ Statistics and visualization
- ✓ Web mining
  - Clustering, classification, and outlier detection
  - Association rule mining and sequential pattern mining
  - Text mining

These categories are further specified:

- ✓ Prediction – develop a model for deducing some aspect of the data from some combination of other aspects of the data (classification, regression, density estimation)
- ✓ Structure Discovery – find structure and patterns in the data sets (clustering, factor analysis, domain structure discovery, network analysis)
- ✓ Relationships – discover relationships between variables in a data set (association rule, correlation, sequential pattern, causal data)
- ✓ Distillation of data for human judgment
- ✓ Discovery with models – developed pre-existing or come from other analysis.

The generated data present snap-shots of participants' interactions during educational process. Analysing and visualising these data can give an idea about:

1) *Various statistics* – the most popular resources, resources valued by students; number of downloads of learning resources; external auxiliary forums, wikis, blogs, vlogs, etc.; used tools, patterns of use over time [8]; the number and duration of visits; most searched topics and terms. On that base can be derived summaries and reports, statistical indicators on the learner's interactions (with online learning environments, learner-to-learner, learner-to-teacher) and expose trends about activities (the time a student dedicates to the course, the learners' behaviour and time distribution, the frequency of studying events, patterns of studying activity, etc.). In addition, statistical graphs about educational attempts and mistakes, assignments completion, exam scores, student

progression, etc. [9]; social, cognitive and behavioural aspects of students [10] also can be obtain.

2) *Feedback for supporting instructors* – association rule mining methods reveals hidden relationships among entities in large databases (between each learning-behaviour pattern so that the teacher to stimulate productive learning behaviour) [11], between thinking styles of learners and the effectiveness of the structure of learning environments [12], identifying engaging learning patterns); solution strategies for improving effectiveness of online education systems (improve the organization and design of course resources to achieve adaptation; assigning tasks and homework at different levels of difficulty [13], automatic gathering feedback on the learning progress in order to evaluate educational courses, special analysis of the data from question tests so that to refine the question database); support the analysis of trends and detect essential teaching methods.
3) *Student's modelling* – building personal profiles (cognitive and psychology characteristics, knowledge background, learning style and behaviour, etc.) in order to assist adaptive learning, as well as to support classification of students to achieve effective group learning
4) *Recommendations for students* – providing proper recommendations to the students according to their profiles and consistent with their educational goals, activities and preferences to assist personalisation.
5) *Predicting student performance* – basing on student profile, records for learning interactions and detecting student behaviours.
6) *Planning, scheduling and constructing curriculum* – assists development process of courses and learning resources automatically and supports reuse of existing learning units. Big Data instruments allow automatically constructing concept maps [14].

Other interesting and unexpected deductions and results about the whole learning process can be derived from the huge amount of data stored. The integration of Big Data concepts with the e-learning systems will lead to the realization of the above listed goals which will improve the effectiveness of the education system**.**

3 Big Data Implementation in Education Sector

One of the reasons for implementation of Big Data techniques in education is that they allow usage of statistical methods which can assist education analytics and decision-makers in identifying possible problems in educational process as a whole (macro context) and responding accordingly in order to raise the effectiveness of educational institutions. On the

other hand, Big Data techniques involve methods and tools for exploring the educational data that allow better understanding of students' needs and requirements in order to improve students' performance by providing personalised recommendations and course content. This type of analysis is focused on the individual characteristics of each student and on the details of learning course's resources. These methods have exploratory aspect, so they can be used for prediction of students' performance and for mapping out strategy for future institutional enhancement. In addition, those approaches can be used for students' modelling and clustering (grouping) in order to provide adapted individual or group learning.

Implementing Big Data in education gives powerful tools to educators to achieve evidence-based teaching/learning process, to apply more flexible, more adaptable and hence personalised approaches, "to go well beyond the scope of what they may want to do" [15]. The successful implementation of Big Data techniques depends on the availability and reliability of education-related information and applied data warehousing strategy where the educational context has to be taken into account, especially corresponding semantic information. According to us, Big Data are especially suitable and of a great importance for managing two kinds of educational information:

1) Information on the learning path:
   - Records of all actions made by the learner regarding the learning process:
     - ✓ Manner of solving a task (approach to the problem)
     - ✓ Speed, quality and scope of the solution
     - ✓ Try outs / errors / needed help (tips, additional information, etc.)
     - ✓ Performance statistics:
       - Ongoing data (comparison with the other learners, automatically generated data)
       - Collected and accumulated data to be analyzed and processed according to certain criteria
   - Record of student's (portfolio)
     - ✓ A priori information about the learner
     - ✓ Chronology of education
     - ✓ Keeping track of learning performance
     - ✓ Statistics about the student's development (selection of periods for analysis, average values of indicators, determination of values associated with applied teaching methods, correlations etc.)
2) Information on the learning resources:

- ✓ Ready for use learning units
- ✓ Means and tools for adapting learning units to individual requirements, problems, situations – adaptive learning systems, personalized learning scenarios (constructing and presenting learning material in various ways)
- ✓ Libraries and repositories with recommended learning resources, not included in the curriculum
- ✓ Other additional resources - popular science, wikis, thesaurus, etc.

To visualise how the above described types of information can impact the education system effectiveness, the Big Data concept should be integrated with the e-learning systems and put into practice. An example scenario for generation, accumulation and interaction of Big Data in educational context that are the base for constructing personalized learning for the student is given on the figure 1.

Using the great facilities provided by Big Data many relations, correlations, dependences, diverse trends and processes can be extracted and visualized, which leads to new insights and new knowledge. They will be directly related to students' learning behaviour and corresponding patterns; can help to identify at-risk students and help institutions to make appropriate prevention [16]. As a consequence educators could take relevant action to support educational processes, which will improve the overall efficiency of the institution and of learning in particular. The applied techniques to reveal the factors for students' success are classification and regression trees that are specific to data mining, also both quantitative and qualitative research and analysis [16]. This research is an example of the successful application of data analysis for avoiding drop-out of students. Although, these results may not be generalized, the applied data techniques can be generalized and reused in similar contexts, but an issue of standardization of data and models should be considered.

Big Data can also be used for assessment and analysis of students' achievements and to predict students' progress in the next grade [17]. The research used a mixed-methods approach – quantitative and case study analysis, which gives the ability to assess a specific educational process. Another often used data mining techniques for the Big Data are association rules, clustering, classification trees, sequential pattern analysis, dependency modelling, multivariate adaptive regression splines, random forests, decision trees and neural networks.

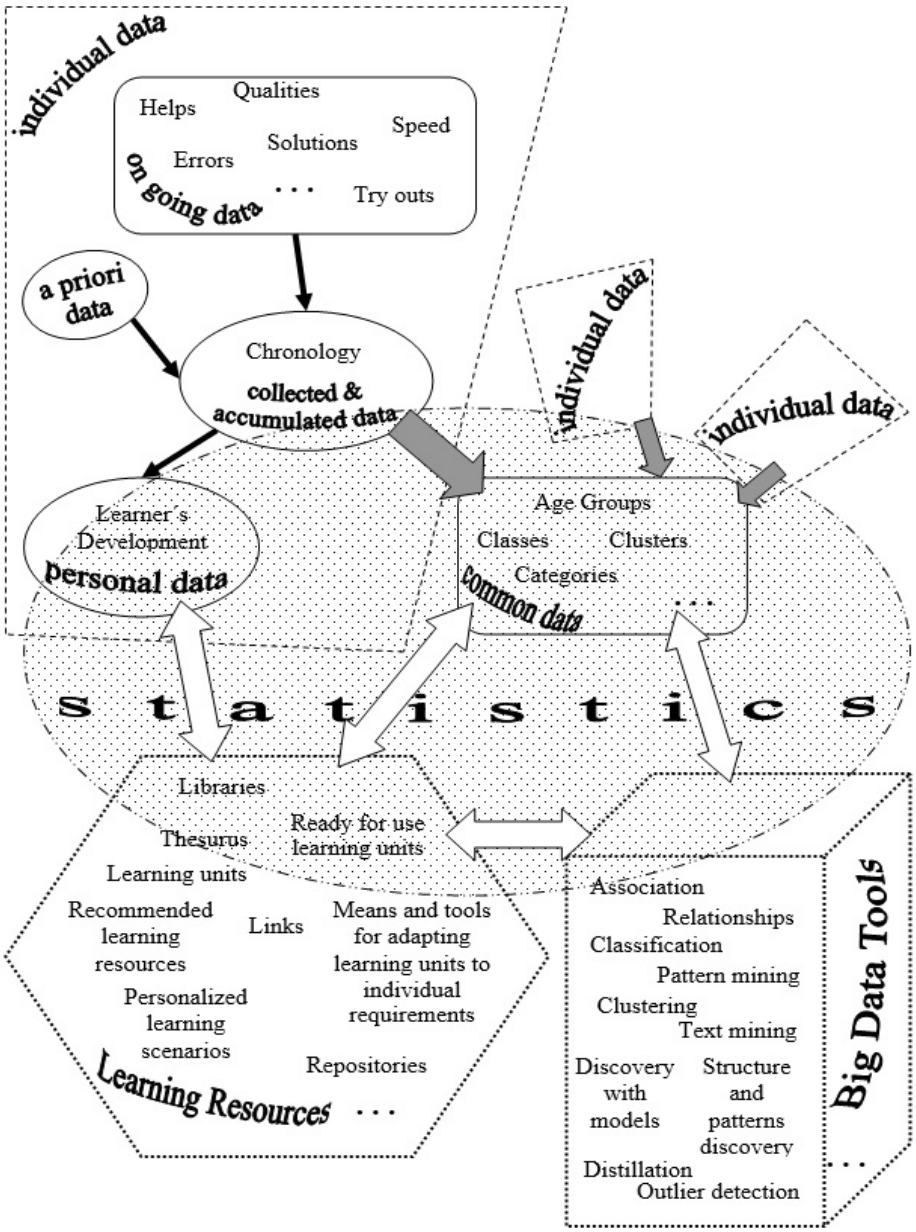


Figure 1. Generation, accumulation and interaction of Big Data in educational context

4 Adaptive and Personalised Learning

Educational Big Data are used also in Personal Learning Environments (PLE) and Personal Recommendation Systems (PRS). PLEs provide tools and services that ensure

instant system adaptation to students' learning needs [18, 19]. The recommendations should coincide with educational objectives so that the system must attempt to understand or determine the needs of learners. Also there should be some way for educators to control recommendations for their students [20].

In e-learning environment students' browsing behaviour play significant role for recommendations for further learning exercises and improves student achievement. The relation between annotated browsing events, contextual factors and access patterns shows that Big Data methods can assist analysis of the individual learner's need and hence delivering highly personalized content, based on records of browsing history (logs/ records of learning path or activities) and students' performance. This method allows students to move through the material at their own pace which also improve student learning.

Another usage of Big Data method is as a way to analyse users' preferences in interactive learning systems. By means of clustering technique students are divided into separate groups according to their preferences and computer experience [21]. Other preferences, such as age, gender, etc. could be used.

Big Data method can be used to provide learners with continuous chain of recommendations to help them learn more effectively and efficiently. A method based on item response theory was used to extract learner behaviour patterns in an online course and subsequently, provide learners with different levels of recommendations rather than single ones [22]. The recommended next piece of learning unit/ resource depends on the performance (answers) on previous. Such system directly impact students' resource selections by providing them with highly individualized recommendations for improved learning efficiency. Students will be provided with help and other options based on their own learning patterns and successful strategies from many other learners who already pass or failed particular learning topic/ resource/ question. Such systems allow educators to adapt content delivery, based on continuous analysis of user experience, with Big Data methods.

Big Data techniques allow by exploring accumulated data for the students' interactions within learning system to be acquired knowledge for correctness of student' responses, time spent on particular task, number of attempts for passing it, number and kind of needed hints/ help, repetitions of wrong answers, and errors made. Such data can be used for creation of learner's skills and knowledge model that can be made automatically by a predictive computer modelling or by a educator. Such models usually are used to customize and adapt the system's behaviours to users' specific needs and preferences so that the systems "say" the "right" thing at the "right" time in the "right" way [23]. These modelling

techniques are widely used in adaptive hypermedia, recommender systems, and intelligent tutoring systems, where knowledge models determine next step in learning path of the students.

Students' behaviour modelling often characterizes their actions and can be used as a clue for their engagement. It can be inferred from the same kinds of data used in learner's skills and knowledge modelling added to some other concerning user behaviour modelling, such as time spent within the system, speed of passing learning resource/ course, number and type of completed courses, regularity of attendance, standardized test scores, etc. Such models help teachers to understand and distinguish student learning behaviour and provide them more engaging teaching.

Usually adaptive learning systems are based on students' knowledge and behaviour modelling in order to provide customized feedback that gives recommendations based on analysis of collected data. Such system uses analytics to deliver only the proper content for the students and skips topics that they have already passed successfully.

Some education experts think that it is possible the individual learning path to be completely data driven. By tracking a student's mastery of each subject or skill, an e-learning system can give just the right piece of instruction resource. Other experts are sceptical about allowing completely automatic determination of knowledge or skills that students have to acquire further or topics to practice next.

Educational Big Data can operate within the e-learning systems to improve student academic outcomes. Non-expert users are allowed to get Big Data information for their courses and teachers are allowed to collaborate with each other and share results [24].

Course management systems can also be mined for usage data to find specific patterns and trends in student online behaviour. Usage data contain information about learner's activities, such as testing, quizzes, reading, comments and discussions, etc. Educational data can be used in order to customise learning path and related activities for individual or group of students. Instead of having static course content, the course is adapted in accordance with student's profile, offering him/ her personalised learning path and his/ her own pace. Learning resources and tasks are also adapted according to students' progress through the course. Usually students begin the course with varying levels of competency. The Big Data method allows to create meaningful optimal learning path for each student [25].

In online educational systems an important factor for learners' success is their engagement with the course content. Big Data methods can be used to determine if there are

disengaged learners by to analysing the speed at which students read through the pages and the length of time spent on pages [26].

Big data can be used in online learning systems to prevent student from manipulating and outwitting the system by cheating and abusing (e.g. clicking until the system provides a correct answer or regular usage of help) in order to make progress, while avoiding learning [27]. So, various modifications of the system that can prevent those issues are provided. These include providing additional exercises and tasks, or using an intelligent agent showing disapproval when detect gaming behaviour.

5 Conclusions

The aim of our paper is to reveal a number of reasons why and how Big Data may revolutionize the education sector. The Big Data techniquie allow convertion of diverce raw data from the whole educational systems into useful information that has the power to impact their working. They provide educational institutions with efficient and effective ways to enhance their effectiveness and students' learning, as well as with useful tools assisting organizations in decision making based on the analysis of relationships and patterns among huge data sets.

There is a variety of benefits that the employment of big data in education can offer to society and individuals. Most of the universities are applying learning analytics in order to improve the provided services, setting policies, and professional qualification. Schools are also starting to adopt such institution-level analyses for detecting problem aspects in order to improve measuring indicators such as grades. Making visible students' learning activities enables students to develop skills in monitoring their own learning and to see directly results of their efforts. Teachers gain views into students' performance that help them adapt their teaching or initiate appropriate interventions such as tutoring, tailored assignments, etc. Educators are able to see on the fly the effectiveness of their adaptations and recommendations, providing feedback for continuous improvements. The online resources enable teaching to be always on hand, educational data management and learning analytics enable learning to be easy assessed. Educators at all levels can benefit from understanding the possibilities of the developed tools using Big Data, which in turn can help to increase the quality and effectiveness of learning and teaching.

References

[1] Briggs S. (2014) Big Data in Education: Big Potential or big Mistake? http://www.opencolleges.edu.au/informed/features/big-data-big-potential-or-big-mistake/

[2] Bellaachia, A., Vommina, E. (2006) MINEL: A Framework for Mining E-Learning Logs. In Vth IASTED International Conference on Web-based Education, Mexico, 259-263.

[3] Romero, C., Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art, IEEE Transactions on Systems, Man, and Cybernetics – Part C, 40(6), 601-618.

[4] Jindal R., Dutta Borah, M. (2013) A Survey on Educational Data Mining and Research Trends International Journal of Database Management Systems, Vol. 5, No. 3, pp 53-73.

[5] Dunham, M. (2003). Data Mining: Introductory and Advanced Topics. Upper Saddle River, NJ: Pearson Education

[6] Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1 (1), 3-16.

[7] Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368-384.

[8] Ingram, A. (1999). Using web server logs in evaluating instructional web sites. In Journal of Educational Technology Systems, 28, 2, 137–157.

[9] Shen, R., Yang, F., Han, P. (2002) Data analysis centre based on e-learning platform. In Workshop The Internet Challenge: Technology and Applications, Berlin, 19–28.

[10] Mazza, R., Vania, D. (2003) The design of a Course Data Visualizator: an Empirical Study. In International Conference on New Educational Environments, Lucerne, 215-220.

[11] Yu, P., Own, C., Lin, L. (2001) On learning behaviour analysis of web based interactive environment, International Conference on Computer and Electrical Engineering, Oslo, 1-9.

[12] Ha, S., Bae, S., Park, S. (2000) Web Mining for Distance Education. In IEEE International Conference on Management of Innovation and Technology, Singapore, 715–719.

[13] Shen, R., Han, P., Yang, F., Yang, Q., Huang, J. (2003) Data mining and case-based reasoning for distance learning. Journal of Distance Education Technologies, 1, 3, 46–58.

[14] Chen, N. S., Kinshuk, Wei, C. W., Chen, H. J. (2008) Mining e-learning domain concept map from academic articles. In: Computers & Education Journal, 50, 1009-1021.

[15] Bari, M. Lavoie, B. (2007) Predicting interactive properties by mining educational multimedia presentations. In International Conference on Information and Communications Technology, 231-234.

[16] Luan, J. (2002) Data Mining and Knowledge Management in Higher Education - Potential Applications. Annual Forum of the Association for Institutional Research, Toronto, Canada.

[17] http://eric.ed.gov/ERICWebPortal/detail?accno=ED474143

[18] Yeats, R., Reddy, P. J., Wheeler, A., Senior, C., & Murray, J. (2010). What a difference a writing centre makes: a small scale study. Education & Training, 52 (6/7), 499-507.

[19] Mödritscher, F. (2010) Towards a recommender strategy for personal learning environments. Procedia Computer Science, 1 (2), 2775-2782 [

[20] Ivanova T., Terzieva V. (2011) Personal learning environment (PLE) - training systems and development environments for customized eLearning Proceedings of IVth National Conference Education in the Information Society, Plovdiv, 249- 258. (in Bulgarian)

[21] Santos, O. C., & Boticario, J. G. (2010) Modelling recommendations for the educational domain. Procedia Computer Science, 1 (2), 2793-2800.

[22] Chrysostomou, K., Chen, S. Y., & Liu, X. (2009) Investigation of Users' Preferences in Interactive Multimedia Learning Systems: A Data Mining Approach. Interactive Learning Environments, 17 (2), 151-163.

[23] Huang, Y.-M., Chen, J.-N., Cheng, S.-C. (2007) A Method of Cross-Level Frequent Pattern Mining for Web-Based Instruction. Educational Technology & Society, 10 (3), 305-319. In: IBM (2012). What is big data? from http://www-01.ibm.com/software/data/bigdata

[24] Gerhard, F. (2001) User Modelling in Human-Computer Interaction. User Modelling and User-Adapted Interaction 11, 65–86.

[25] García, E., Romero, C., Ventura, S., de Castro, C. (2011) A collaborative educational association rule mining tool. The Internet and Higher Education, 14 (2), 77-88.

[26] Wang, Y.-h., & Liao, H.-C. (2011) Data mining for adaptive learning in a TESL-based e- learning system. Expert Systems with Applications, 38 (6), 6480-6485.

[27] Cocea, M., & Weibelzahl, S. (2009) Log file analysis for disengagement detection in e-Learning environments. User Modelling and User - Adapted Interaction, 19 (4), 341-385.

[28] Muldner, K., Burleson, W., Van de Sande, B., & Vanlehn, K. (2011) An analysis of students' gaming behaviours in an intelligent tutoring system: predictors and impacts. User Modelling and User - Adapted Interaction, 21 (1-2), 99-135.

Comparison of Two Kinds of Cooperation of Substantial Agents

František Čapkovič*, Lyubka Doukovska**, Vassia Atanassova**

*Institute of Informatics, Slovak Academy of Sciences,

Dúbravská cesta 9, 845 07 Bratislava, Slovakia

Frantisek.Capkovic@savba.sk

**Institute of Information and Communication Technologies, Bulgarian Academy of Sciences,

Acad. G. Bonchev Str., Block 2, 1113 - Sofia, Bulgaria

l.doukovska@mail.bg, vassia.atanassova@gmail.com

Abstract: Place/transition Petri nets (P/T PN) are used here in order to model the behaviour of agents as well as the agent communication which is necessary for their cooperation and negotiation. As the agents, the substantial devices (e.g. industrial robots) are understood. Two kinds of the agent communication will be compared. The first kind of the communication assumes that all agents are equal, i.e. no agent has a higher priority than others. The second kind of the communication has the hierarchical structure. Here, a supervisor on the upper level hsa higher priority than other agents. It coordinates activities of the agents being on the lower level. Here, the individual agents do not cooperate directly, but through a supervisor.

Keywords: Agent, communication, cooperation, hierarchy, negotiation, robot, supervisor.

1. Introduction and Preliminaries

Because the Petri nets (PN) will be used for modeling both the agents behavior and the communication among them, it is necessary to introduce PN. As to the structure place/transition PN (P/T PN) are bipartite directed graphs, i.e. the digraphs with two kinds of nodes (places and transitions) and two kind of edges (from places to transitions and from transitions to places). Thus PN structure is given formally as follows

$$< P, T, F, G >, \quad P \cap T = O, \quad F \cap G = O \tag{1}$$

where P is the set of the PN places p_i , i = 1, 2, …, n; T is the set of the PN transitions t_j, j = 1, 2, …, m; $F \subseteq P \times T$ is the set of edges directed from the places to the transitions; $G \subseteq T \times P$ is the set of edges directed from the transitions to the places; O is the empty set.

However, PN have also their dynamics – the marking evolution of the places. Formally, the dynamics can be expressed by the following quadruplet

$$< X, U, \delta, \mathrm{x}_0 >, \quad X \cap U = O \tag{2}$$

where X is the set of the PN state vectors $\mathbf{x}_k$ and U is the set of the control vectors $\mathbf{u}_k$ of the PN with $k=0, 1, \cdots, N$ being the step of the dynamics evolution; $\delta: X \times U \to X$ is the PN transition function expressing formally that a new state is given on the base of existing state and the occurrence of discrete events; $\mathbf{x}_0$ is the initial state vector of the PN.

The system form of the PN-based model of a DES module is the following

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{B}.\mathbf{u}_k, \quad k = 0,1,\cdots,N \tag{3}$$

$$\mathbf{B} = \mathbf{G}^T - \mathbf{F} \tag{4}$$

$$\mathbf{F}.\mathbf{u}_k \le \mathbf{x}_k \tag{5}$$

where k is the discrete step of the dynamics development; $\mathbf{x}_k = (\sigma_{p1}^k, \cdots, \sigma_{pn}^k)^T$ is the n-dimensional state vector; $\sigma_{pi}^k \in \{0,1,\cdots,c_{pi}\}, i = 1,\cdots,n$ express the states of atomic activities of the PN places by 0 (passivity) or by $0 < \sigma_{pi} \le c_{pi}$ (activities); c_{pi} is the capacity of p_i; $\mathbf{u}_k = (\gamma_{t1}^k, \cdots, \gamma_{tm}^k)^T$ is the m-dimensional control vector; $\gamma_{tj}^k \in \{0,1\}, j = 1,\cdots,m$ represent occurring of the elementary discrete events (e.g. starting or ending the activities, failures, etc.) by 1 (presence of the corresponding discrete event) or by 0 (absence of the event); **B**, **F**, **G** are matrices of integers; $\mathbf{F} = \{f_{ij}\}_{n \times m}$, $f_{ij} \in \{0, M_{fij}\}$, expresses the causal relations among the states (being causes) and the discrete events occurring during the DES operation (being consequences) by 0 (nonexistence of the relation) or by $M_{fij} > 0$ (existence and multiplicity of the relation); $\mathbf{G} = \{g_{ij}\}_{m \times n}, g_{ij} \in \{0, M_{gij}\}$, expresses analogically to the previous matrix the causal relations among the discrete events (being causes) and the DES states (being consequences); **B** is given according to the equation (4) and symbolizes the system parameters; $(.)^T$ symbolizes the matrix or vector transposition. It is necessary to say that all parameters and variables are nonnegative integers.

Agents are usually understood to be [3] persistent (software, but not only software) entities that can perceive, reason, and act in their environment and communicate with other agents. From the external point of view the agent (a real or virtual entity): (i) evolves in an environment (ii) is able to perceive this environment; (iii) is able to act in this environment;

(iv) is able to communicate with other agents; (v) exhibits an autonomous behaviour. From the internal point of view the agent encompasses some local control in some of its

perception, communication, knowledge acquisition, reasoning, decision, execution, and action processes.

Here, the substantial agents will be used. Namely, the communication, necessary for the cooperation and negotiation [1], [2] of industrial robots, will be studied. First of all the group of agents, where no agent has a higher priority than other ones, will be modelled by P/T PN and their communication will be analyzed. Then, the hierarchical structure of the agents will be examined. Here, two levels of the hierarchy will be taken into account. On the lower level the agents with the same priority will be placed, without any possibility to communicate each other. On the upper level the agent-supervisor, representing the cooperation strategy, will be placed. This agent will ensure the purposeful communication of the agents tending to the global goal of the whole group of the agents. Finally, both of the structures of the agent communication will be compared.

2. Case Study

Two kinds of the organization structures of the agent group will be introduced and analyzed in this section, namely, the structure with the free communication of the agents each other and the two-level hierarchical structure with the agent-supervisor on the upper level. For simplicity, the group consisting of three substantial agents (industrial robots) will be examined.

2.1 The Structure of the Group of Agents with the Free Communication

Consider three substantial agents - intelligent robots A_1, A_2, A_3. The P/T PN models of them are given in Fig. 1. Each of them has the same structure. The PN models of them are created by means of the sets of PN-places $P_{A1} = \{p_1, p_2, p_3\}$, $P_{A2} = \{p_4, p_5, p_6\}$, $P_{A3} = \{p_7, p_8, p_9\}$ and by means of the sets of PN transitions $T_{A1} = \{t_1, t_2, t_3, t_4\}$, $T_{A2} = \{t_5, t_6, t_7, t_8\}$, $T_{A3} = \{t_9, t_{10}, t_{11}, t_{12}\}$. The PN places represent three basic states of the agents. Namely, they interpretation is the following: the agents are either available (p_2, p_5, p_8) or they want to communicate (p_3, p_6, p_9) or they do not want to communicate (p_1, p_4, p_7). The communication channels between two corresponding agents have the unified structure. The PN model of the channel Ch_1 between A_1 and A_2 consists of $\{(p_{10}, p_{11}), (t_{13}, t_{14}, t_{15}, t_{16})\}$, the model of the channel Ch_2 between A_1 and A_3 consists of $\{(p_{12}, p_{13}), (t_{17}, t_{18}, t_{19}, t_{20})\}$, and the model of the channel Ch_3 between A_2 and A_3 consists of $\{(p_{14}, p_{15}), (t_{21}, t_{22}, t_{23}, t_{24})\}$. The interpretation of the places (states of the channels) are: the channels are either available (p_{11}, p_{13}, p_{15}) or realizing the communication of corresponding agents (p_{10}, p_{12}, p_{14}). The models of the

channels are also given in Fig. 1. The incidence matrices of the PN models of the agents are the following

$$\mathbf{F}_{Ai}=\begin{pmatrix}0&1&0&0\\1&0&1&0\\0&0&0&1\end{pmatrix};\ \mathbf{G}_{Ai}^{T}=\begin{pmatrix}1&0&0&0\\0&1&0&1\\0&0&1&0\end{pmatrix};\ \mathbf{x}_0^{Ai}=\begin{pmatrix}0\\1\\0\end{pmatrix};\ i=1,2,3 \qquad (6)$$

and the incidence matrices of the PN models of the communication channels are as follows

$$\mathbf{F}_{Chi}=\begin{pmatrix}0&0&1&0\\1&1&0&1\end{pmatrix};\ \mathbf{G}_{Chi}^{T}=\begin{pmatrix}1&1&0&1\\0&0&1&0\end{pmatrix};\ \mathbf{x}_0^{Chi}=\begin{pmatrix}0\\1\end{pmatrix};\ i=1,2,3 \qquad (7)$$

The global structure given in Fig. 1 models also the agent communication. The mutual communication (cooperation and/or negotiation) of the agents through the channels is realized by means of firing the transitions. The transition t_{16} is fired when both agents A_1, A_2 want to communicate, t_{14} is fired when A_1 wants to communicate with A_2 and A_2 have no objection, and t_{13} is fired when A_2 wants to communicate with A_1 and A_1 have no objection. Analogically, t_{20} is fired when both agents A_1, A_3 want to communicate, t_{18} is fired when A_1 wants and A_3 have no objection, and t_{17} is fired when A_3 wants and A_1 have no objection as well as t_{24} is fired when both agents A_2, A_3 want to communicate, t_{22} is fired when A_2 wants and A_3 have no objection, and t_{21} is fired when A_3 wants and A_2 have no objection. The communications channels create the interface between the communicating agents. They also can be understood to be the agents. The PN model of the entire group of the communicating agents can be expressed as

$$\mathbf{F}_A=blockdiag\left(\mathbf{F}_{Ai}\right);\ \mathbf{G}_A^{T}=blockdiag\left(\mathbf{G}_{Ai}^{T}\right);\ i=1,2,3 \qquad (8)$$

$$\mathbf{F}_{Ch}=blockdiag\left(\mathbf{F}_{Chi}\right);\ \mathbf{G}_{Ch}^{T}=blockdiag\left(\mathbf{G}_{Chi}^{T}\right);\ i=1,2,3 \qquad (9)$$

$$\mathbf{F}=\begin{pmatrix}\mathbf{F}_A&\mathbf{F}_c\\\mathbf{0}&\mathbf{F}_{Ch}\end{pmatrix};\ \mathbf{G}^{T}=\begin{pmatrix}\mathbf{G}_A^{T}&\mathbf{G}_c^{T}\\\mathbf{0}&\mathbf{G}_{Ch}^{T}\end{pmatrix};\ \mathbf{x}_0=\begin{pmatrix}\mathbf{x}_0^{A}\\\mathbf{x}_0^{Ch}\end{pmatrix} \qquad (10)$$

$$\mathbf{x}_0^{A}=\begin{pmatrix}\mathbf{x}_0^{A1}\\\mathbf{x}_0^{A2}\\\mathbf{x}_0^{A3}\end{pmatrix};\ \mathbf{x}_0^{Ch}=\begin{pmatrix}\mathbf{x}_0^{Ch1}\\\mathbf{x}_0^{Ch2}\\\mathbf{x}_0^{Ch3}\end{pmatrix} \qquad (11)$$

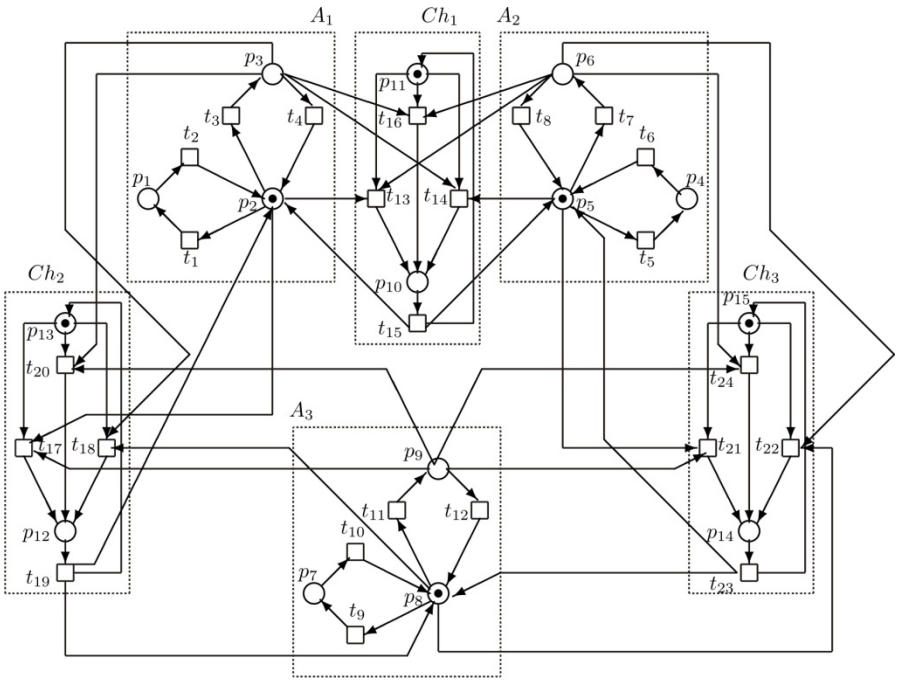


Figure 1. The communication of three agents A_1, A_2, A_3 each other by means of the communication channels Ch_1, Ch_2, Ch_3

$$\mathbf{F}_c = \begin{pmatrix} 0&0&0&0&0&0&0&0&0&0&0&0 \\ 1&0&0&0&1&0&0&0&0&0&0&0 \\ 0&1&0&1&0&1&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&1&0&0&0 \\ 1&0&0&1&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&1&0&0 \\ 0&0&0&0&1&0&0&1&1&0&0&1 \end{pmatrix}; \mathbf{G}_c^T = \begin{pmatrix} 0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0 \end{pmatrix}$$

Having the complete parameters and initial state of the PN model we can simulate the behaviour of the group of the agents. There are two possibilities how to do this: (i) by means a graphical PN editor – such a tool yields also the possibilities for testing the PN model properties (safeness, liveness, boundedness, etc.) and for computing and drawing the reachability tree (RT); (ii) by means of the simulation tool Matlab. Here, the numerical computations can be realized and the adjacency matrix of the RT and the feasible states can be obtained.

The model drawn in the PN editor by means of icons is given in Fig. 2. Starting from the initial state $\mathbf{x}_0$ (see active PN places in Fig. 2) when A_1 and A_2 are going to cooperate, 35 different states are reachable. They create the root of RT (i.e. $\mathbf{x}_0$) and the RT leaves (i.e. all state vectors $\mathbf{x}_k$ reachable from $\mathbf{x}_0$).

As we can see in Fig. 1 and/or in Fig. 2 the structure mentioned and described above needs the communication channel between any two agents. Thus, in case of N agents the number of needful channels is

$$N_{Ch} = \begin{pmatrix} N \\ N-2 \end{pmatrix} \tag{12}$$

i.e. for 4 agents 6 channels are needed, for 5 agents 10 channels, for 6 agents 15 channels, etc. In addition, the RT corresponding to the PN model – yielding the number of the states reachable from the initial state $\mathbf{x}_0$ and the relations among them – is large too.

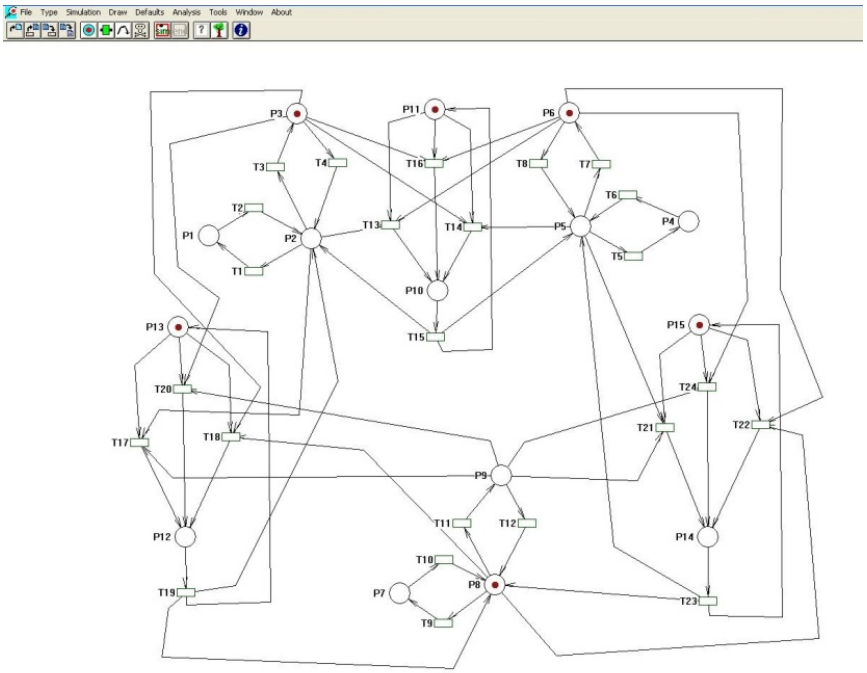


Figure 2. The PN model drawn in the PN editor

In such a case, the process of negotiation can be complicated, primarily in the case of a larger number of agents. Moreover, in industrial applications, especially in manufacturing systems, there is no time for such a circuitous process of cooperation and negotiation. In such a case economics of the production can be deeply damaged. Therefore, a more suitable structure has to be found.

2.2 Hierarchical Structure of the Cooperation and Negotiation

As an alternative to the previous structure, the hierarchical structure is being touted. On the lower level the individual robots will be placed, without any possibility to communicate each other. On the upper level the agent-supervisor, representing the cooperation strategy, will be placed. The supervisor will ensure the purposeful communication of the agents tending to the global goal of the whole group of the agents. Namely, the agent-supervisor enforces a global strategy of the whole group into the particular activities of the individual agents working on the lower level. On the one hand, the autonomy of individual agents is disturbed, but on the other hand, the activity of the entire group is more effective. Consequently, the global goal can be achieved more directly and perhaps also more quickly.

Here, the number of communication channels is equal to the number of agents on the lower level of the hierarchy – i.e. N. There are no interconnections among these agents. For example in the case of $N = 3$ agents, the structure of the communication is given in Fig. 3. However, in such a structure the supervisor is not able to communicate simultaneously with more robots, only with one. The structural matrix of the model is the following

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_S & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_1 & \mathbf{X}_1 & \mathbf{X}_1 \\ \mathbf{0} & \mathbf{B}_{A1} & \mathbf{0} & \mathbf{0} & \mathbf{X}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{A2} & \mathbf{0} & \mathbf{0} & \mathbf{X}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{A3} & \mathbf{0} & \mathbf{0} & \mathbf{X}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{Ch1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{Ch2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{Ch3} \end{pmatrix} \tag{13}$$

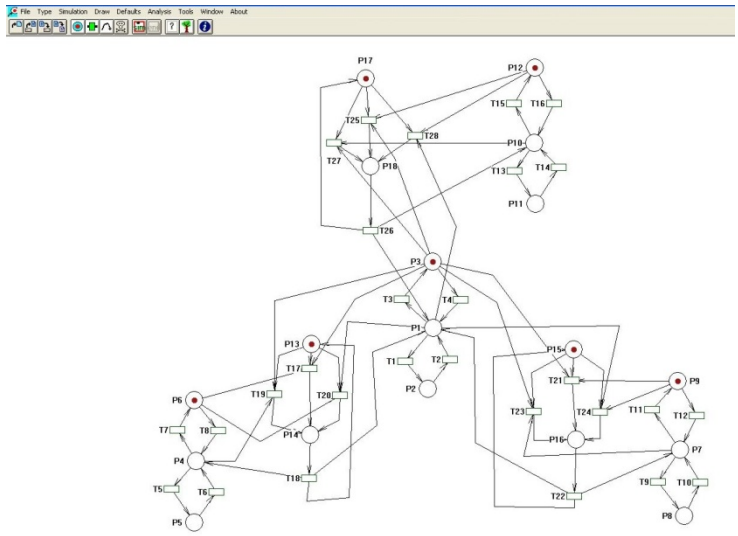


Figure 3. The hierarchical structure of the agent communication

where the structural matrices of the supervisor, agents and channels are as follows

$$\mathbf{B}_S = \begin{pmatrix} -1 & 1 & -1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix};\ \mathbf{B}_{Ai} = \mathbf{B}_S,\ i = 1,2,3;\ \mathbf{B}_{Chi} = \begin{pmatrix} -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 \end{pmatrix},\ i = 1,2,3$$

While the structural matrices of the interconnections are

$$\mathbf{X}_1 = \begin{pmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 \end{pmatrix};\ \mathbf{X}_2 = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 \end{pmatrix}$$

The structure of the state vector has the form

$$\mathbf{x}_0 = \left((\mathbf{x}_0^S)^T \quad (\mathbf{x}_0^{A1})^T \quad (\mathbf{x}_0^{A2})^T \quad (\mathbf{x}_0^{A3})^T \quad (\mathbf{x}_0^{Ch1})^T \quad (\mathbf{x}_0^{Ch2})^T \quad (\mathbf{x}_0^{Ch3})^T\right)^T \tag{14}$$

For the initial state $\mathbf{x}_0 = (0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0)^T$ the system has 107 states reachable from $\mathbf{x}_0$.

2.3 The Comparison of the Structures

Now, compare here the presented structures. Each of the structures has its advantages and disadvantages. The main advantage of the former communication structure is that the agents can communicate each other directly, without any intermediary. Paradoxically, simultaneously it is also a disadvantage. Namely, the communication facility has to be placed

between any pair of the agents. Each agent must be able to communicate with any other agent. Consequently, a big number of communication devices must be used. Such a disproportion can be resolved using parallel control. Moreover, the communication of an agent with more than one partner has to be eliminated. Really, each agent is allowed to communicate simultaneously only with one partner. The main advantage of the latter communication structure is the smaller number of the communication facilities. Their number is equal to the number of agents. However, although on the one hand it is the advantage, on the other hand it is also a disadvantage, because the agents are not able to communicate each other directly but only by means of the supervisor being the intermediary. The supervisor is allowed to communicate only with one of the agents simultaneously.

Because we are interested here in the material agents (robots), it is necessary to look at the thing from this point of view. In the flexible production systems, just the second kind of the communication structure is preferable. Indeed, from the economic point of view, there is no time for a lengthy communication associated with the cooperation and negotiation of material agents. The activities of the material agents have to correspond with the technological process being used at the production (i.e. the sequence of operations to be performed) which is the most important factor of the production systems. The agents must not drain in the long lasting mutual fights. Moreover, in this case the PN model is not so large like in the case of the first kind of the communication. Consequently, also the simulation by means of the graphical tool is simpler. However, the smaller model is not any guarantee of a small reachability tree representing the number of the states of the system and their causal interconnections.

3. Conclusions

Two structures of the communication among the material agents (like robots) were studied. Namely, the structure where the agents can communicate each other directly and the structure where their communication is mediated by the supervisor. The behaviour of the agents as well as the communication channels were modeled by P/T PN. The complete PN models of both structures were built. The advantageous and disadvantageous of both structures were compared and evaluated.

References

[1] Čapkovič, F. (2007) Modelling, Analysing and Control of Interactions Among Agents in MAS. *Computing and Informatics*, 26, 5, Slovak Academy of Sciences, 507-541.

[2] Čapkovič, F. (2014) Cooperation of Agents in Complex Systems Based on Supervision. *Cybernetics and Information Technologies*, 14, 1, Bulgarian Academy of Sciences, 40-51.

[3] Saint-Voirin, D., Lang, C., Guyennet, H., Zerhouni, N. (2007) Scoop Methodology: Modeling, Simulation and Analysis for Cooperative Systems. *Journal of Software*, 2, 4, Academy Publisher Inc. UK, 32–42.

Big Data Platform for Monitoring Indoor Working Conditions and Outdoor Environment

Igor Mishkovski

Faculty of Computer Sciences and Engineering, P.O. Box 393, 1000 Skopje, R. Macedonia.

e-mail: igor.miskovski@finki.ukim.mk

Lasko Basnarkov

Faculty of Computer Sciences and Engineering, P.O. Box 393, 1000 Skopje, R. Macedonia.

e-mail: lasko.basnarkov@finki.ukim.mk

Ljupcho Kocarev

Macedonian Academy of Sciences and Arts. Bul. Krste Misirkov 2, P.O. Box 428.

e-mail: lkocarev@manu.edu.mk

Svetozar Ilchev

Institute of Information and Communication Technologies – BAS, Acad. G. Bonchev Str., Bl. 2, 1113 Sofia, Bulgaria

e-mail: svetozar@ilchev.net

Rumen Andreev

Institute of Information and Communication Technologies – BAS, Acad. G. Bonchev Str., Bl. 2, 1113 Sofia, Bulgaria

e-mail: rumen@isdip.bas.bg

Abstract: This paper presents a framework and a test bed for monitoring indoor working conditions and outdoor environment. In particular, we focus on developing: i) a test bed small Wireless Sensor Network (WSN) for monitoring indoor and outdoor environment parameters using low cost Arduino controllers, transceivers and sensors, and ii) real-time platform for analysis and mining of the sensor data. Together the two integral parts will offer not only monitoring, but also mining of the data and detection of any environmental anomalies.

Keywords: Big Data, Sensor, WSN, Mining

1. Introduction

The environment changes quickly, and these changes influence the citizens' health, perceived quality of life and work efficiency. Environmental changes also influence the economy directly e.g. tourism and agriculture. Thus, a great part of the research community is still searching for the right kind and amount of data and analysis tools necessary to address

serious problems that occur unexpectedly and develop rapidly. Furrthermore, the devices with sensing capabilities are becoming ubiquitous, e.g. low-power sensor networks or mobile and wearable devices equipped with sensors. On the other hand, data mining, machine learning are now able to deal with large-scale data sets that contain millions of high-dimensional data points.

The proper monitoring of the outdoor environment and alerting when certain anomalies arise, addresses the major environmental health treat [1]-[5]. However, outdoor environment monitoring, will not only influence the public health, but also, the quality of life and the working efficiency of the citizens.

On the other hand, monitoring the indoor working conditions can improve health, work performance and school performance, reduce health care costs and be a source of substantial economic benefit [6]-[11].

The collected data, both outdoor and indoor, can be collected in data storage and used together with the real-time stream of data for visualizing, anomaly detection and building a model for prediction of employee productivity, etc. In order to process all the sensor information, aggregate it and disseminate it back to the users in relevant way a central Real-Data Processing System will be required, as in [12]. This system will queue and map heterogeneous data and will serve as a service real-time layer for different Distributed Remote Procedure Calls (DRPC). Besides offering different open web services, the platform can offer semantically annotated linked data.

The test bed Wireless Sensor Network (WSN) will statically monitor the environment parameters, as well as it will address mobile monitoring of the condition in the outdoor environment. The collected data will be stored on the Real-Data Processing System, from which we will do indoor and outdoor layered visualization on the building plan and/or GIS systems, respectively. Using the stored data and the real-time data the system will infer anomalies, as well as, using a measure of the employee productivity it will learn which are the conditions that increase the employees' productivity.

Thus, in this work we propose a simple, open and cheap framework that offers data services, both raw and processed. This data allows modelling and exploration of the relations between variables in environment and detection of alarming trends. The platform also will provide feedback to businesses and citizens about influence of their actions on the environment.

The paper is organized as follows. In Section 2 we give the overview of the proposed sensing framework and the possible sensors that could be used for monitoring indoor

working conditions and outdoor environment. Section 3 gives overview of the Real-Data Processing System and Section 4 concludes this paper.

2. Architecture for Indoor and Outdoor Monitoring

The diagram in Figure 1 shows the possible architecture for indoor and outdoor monitoring. The architecture is consisted of several Arduino static nodes that will monitor the indoor conditions. In the diagram, the architecture will measure several parameters, such as Pressure (P), Temperature (T), Humidity (H), Dust and data from other possible sensors. The sensor nodes will continuously read the status of all attached sensors and pass the sensor data through the radio network back to the gateway. These sensors will have the option to sleep most of the time in order to save battery. However, in the system there might exist repeater-sensor nodes (not shown in Figure 1) which must stay awake in order to pass messages from their child sensor nodes. A repeater-node can optionally include direct-attached sensors and report their sensor data to the gateway.

The Arduino sensor nodes will communicate with the Arduino Gateway using the NRF24L01+ transceiver from Nordic Semiconductors which communicates with the Arduino board via the SPI interface. The Arduino Gateway on the other hand will act as a glue between the controller and the radio network. It will translate radio messages to a protocol which can be understood by a controller. There are several possible implementations for the gateway:

- SerialGateway - The gateway connects directly to the controller using one of the available USB ports.
- EthernetGateway - The gateway connects to the Ethernet network that the controller also uses offering more placement flexibility than the SerialGateway.
- MQTTGateway - This gateway also connects to the Ethernet network and exposes an MQTT broker which can be used for controllers offering MQTT support like OpenHAB [13].

AS a good candidate for the controller we will develop our own simple DIY cloud-enabled gateway controller running on the Raspberry Pi.

Besides the serial communication between the controller and the Arduino gateway the controller will collect FTP Data sent by the GPRS module of the mobile Arduino sensor nodes. The mobile Arduino sensor nodes, besides the GPRS module will be equipped with additional sensors for P, T, H, Shinyei PPD42NS Particle sensor, GPS sensor, and some other possible sensors. All the sensor nodes will be boxed in special cases using 3D printer.

The controller will collect all the indoor and outdoor sensor data through the serial and ftp communication, respectively and will feed the data into the real-data processing system.

2.1. Other Relevant Sensor Data and Sensing the Employees' Productivity

Beside the abovementioned sensors, the system can be upgraded with additional sensors that can fed more data into the real data processing system, such as:

- Open/closed doors or the state of a wall switch.
- Distance sensor - it can measure the sitting habits of an employee.
- Gas sensor - for detecting alcohol, methane, fire, etc.
- Infrared sensors - that can control the air-conditioning system in the office.
- Light sensor - can be used in the automated control of the drapers.
- Movement sensor - to detect if the office is overcrowded and what are the dynamics in the office.
- Relay Actuator - to turn on/off the devices.
- RFID sensors – to detect the workers in the office.
- Infrared sensor.
- Noise meter – to measure the noise conditions in the office.
- UV sensor – to measure the UV factor in the environment.

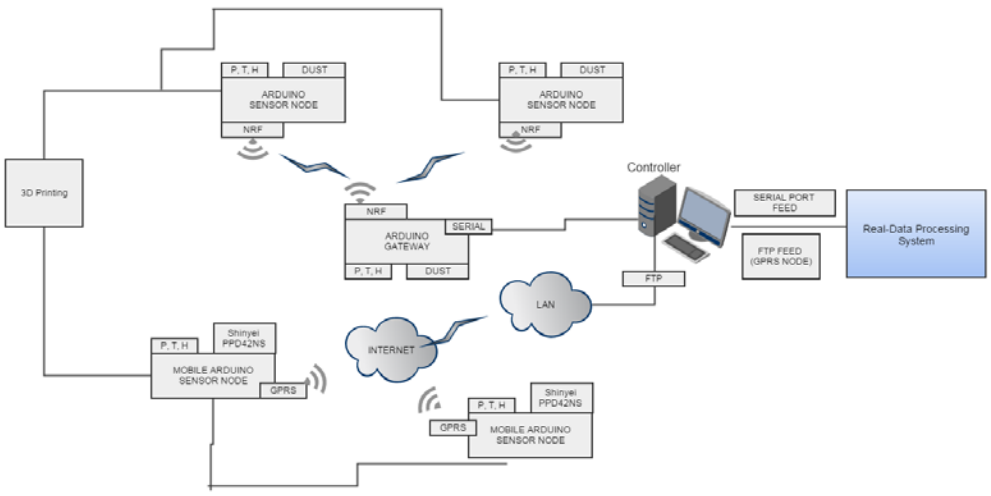


Figure 5 Architecture for Indoor and Outdoor Monitoring

In order to measure the employee productivity we propose four distinct way that will be used as an output feature of the machine learning techniques, such as:

- Using user input – with the help of a button input (productive/non-productive day)
- Measure productivity by gathering the production data, i.e. how many pieces or products were produced in one day in some factory
- Sensor on chairs – using this sensing data we will obtain information how much and what are the sitting habits of the employee.
- Use the data from some corporate task platform (such as google task, or some proprietary).

3. Real Data Processing System

The real-data processing system will be fed by the controller and it provides an architectural model that scales and which has both the advantages of long-term batch processing and the freshness of a real-time system, with data updated in seconds' time.

The data will be are fed into the system (1), for example through a queue from where the system (such as Storm) can pull them. A system, such as Trident, will save them into Hadoop (HDFS) and processes them in real-time for creating an in-memory state. In Hadoop all the historical data will be available, and in any time a batch process can be started that will aggregate the data and generate a big file from it. After that, we can use some API tools or SQL command line tools (such as Splout) to index the file and deploy it to a Splout SQL cluster (4), which will be able to serve all the statistics pretty fast. Then, a second stream (DRPC), such as Trident, can be used to serve timeline queries, and this stream will query both the batch layer (through Splout SQL) and the real-time layer (through the first stream's memory state), and mix the results into a single timeline response. In this way, we will prepare both the historical and the real data for the stakeholders in order to visualize it, receive alerts, and do some possible real-time optimizations.

4. Conclusions

The proposed platform, based on data collected from various sources and processed by online services, should help decision makers in finding balance, optimal trade-offs in real-time. Moreover, using this simple and cheap platform, the stakeholders not only that can obtain various information about the indoor and the outdoor conditions, they can create a model for the employee productivity depending on the conditions and certain anomalies in environment that affect the production. Finally, the employers can affect some of the indoor conditions in order to boost the employee productivity using the obtained models.

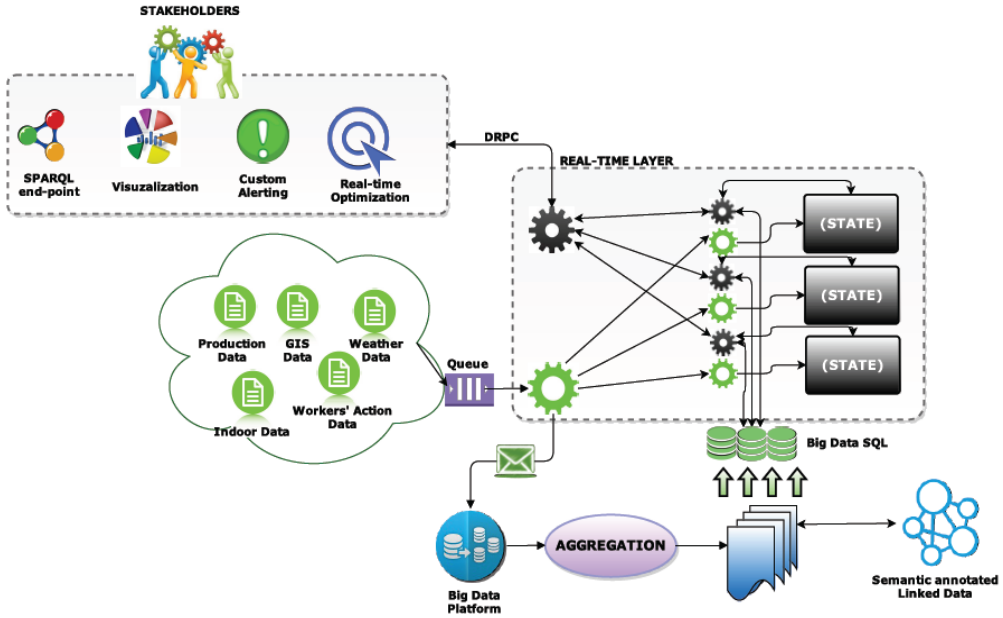


Figure 6 Real-data processing system

References

[1] Kampa, M. & Castanas, E. 2008. Human health effects of air pollution, Environmental Pollution 151:362-367.

[2] Brunekreef, A. & Holgate, S.T. 2002. Air pollution and health. Lancet 360:1233-1242.

[3] Dockerty, D.W., Arden Pope, C., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G. & Speizer, F.E. 1993. An Association Between Air Pollution And Mortality In Six U.S. Cities. The New England Journal of Medicine 329(24):1753- 1759.

[4] Pope C. A. III, Burnett, R.T., Thun, M. J., Calle, E.E., Krewski, D., Ito, K. & Thurston, G. D. 2002. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. Journal of the American Medical Association 287:1132-1141.

[5] Mucke, H.-G. 2000. Ambient air quality programmes for health impact assessment in the WHO European region, Arh Hig Rada Toksikol 51:257-564.

[6] ASHRAE(2011) ASHRAE Position Document on Indoor Air Quality, Atlanta, GA, USA, American Society of Heating, Refrigeration, and Air-Conditioning Engineers, Inc.

[7] Schell, M. and Inthout D. 2001. Demand Controlled Ventilation Using CO2, ASHRAE Journal.

[8] Stranger, M., Potgieter-Vermaak, S.S., Van Grieken, R. 2008. Characterization of indoor air quality in primary schools in Antwerp, Belgium. Indoor Air 2008; 18:454- 463.

[9] Currie, J., Hanushek, E.A., Kahn, E.M., Neidell, M., and Rivkin, G. 2009. Does Pollution Increase School Absences? The Review of Economics and Statistics, November 2009, 91(4): 682-694.

[10] http://www.cdc.gov/HealthyYouth/asthma/

[11] Maclean, M.; Anderson, J.G.; MacGregor, S.J.; Mackersie, J.W. 2004. "The development of a pulsed UV-light air disinfection system and its application in university lecture theatres," Power Modulator Symposium, 2004 and 2004 High-Voltage Workshop. Conference Record of the Twenty-Sixth International , vol., no., pp. 630- 633, 23-26 May 2004.

[12] Rizea, Daniel-Octavian, Olteanu, Alexandru-Corneliu, Tudose, Dan-Stefan. 2014. “Air quality data collection and processing platform”, RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference, Moldova, September 2014.

[13] openHAB, http://www.openhab.org/.

Prinded:

(Info)
(Info)
(Info)
(Info)
(Info)
(Info)
(Info)

ISDN / ISSN:……

Institute of Information and Communication Technologies
- Bulgarian Academy of Sciences

John Atanasoff Society of Automatics and Informatics