THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

# DATABASE SYSTEMS JOURNAL

Vol. VII, Issue 1/2016

LISTED IN

RePEc, EBSCO, DOAJ, Open J-Gate,
Cabell's Directories of Publishing Opportunities,
Index Copernicus, Google Scholar,
Directory of Science, Cite Factor,
Electronic Journals Library

BUSINESS INTELLIGENCE

ERP

DATA MINING

DATA WAREHOUSE

DATABASE

ISSN: 2069 – 3230
dbjournal.ro

### Contact
Calea Dorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: http://dbjournal.ro
E-mail: editordbjournal@gmail.com; editor@dbjournal.ro

# CONTENTS

# Tuning I/O Subsystem: A Key Component in RDBMS Performance Tuning

Hitesh Kumar SHARMA[1], Christalin NELSON. S[2], Dr. Sanjeev Kumar SINGH[3]
[1]Assistant Professor (SS), University of Petroleum & Energy Studies
[2]Assistant Professor (SG), University of Petroleum & Energy Studies
[3]Associate Professor, Galgotia University Noida

hkshitesh@gmail.com, cnelson@ddn.upes.ac.in, sksingh8@gmail.com

**Abstract:** *In a computer system, the fastest storage component is the CPU cache, followed by the system memory. I/O to disk is thousands of times slower than an access to memory. This fact is the key for why you try to make effective use of memory whenever possible and defer I/Os whenever you can. The majority of the user response time is actually spent waiting for a disk I/O to occur. By making good use of caches in memory and reducing I/O overhead, you can optimize performance. The goal is to retrieve data from memory whenever you can and to use the CPU for other activities whenever you have to wait for I/Os. This paper examines ways to optimize the performance of the system by taking advantage of caching and effective use of the system's CPUs.*

**Keywords:** *Tuning, I/O, RDBMS.*

## 1 Introduction

I/O is probably one of the most common problems facing RDBMS users. In many cases, the performance of the system is entirely limited by disk I/O. In some cases, the system actually becomes idle waiting for disk requests to complete. We say that these systems are *I/O bound* or *disk bound*. Disks have certain inherent limitations that cannot be overcome. Therefore, the way to deal with disk I/O issues is to understand the limitations of the disks and design your system with these limitations in mind. Knowing the performance characteristics of your disks can help you in the design stage. Optimizing your system for I/O should happen during the design stage. Different types of systems have different I/O patterns and require different I/O designs. Once the system is built, you should first tune for memory and then tune for disk I/O. The reason you tune in this order is to make sure that you are not dealing with excessive cache misses, which cause additional I/Os. The strategy for tuning disk I/O is to keep all drives within their physical limits. Doing so reduces queuing time and thus increases performance. In your system, you may find that some disks process many more I/Os per second than other disks. These disks are called "hot spots." Try to reduce hot spots whenever possible. Hot spots occur whenever there is a lot of contention on a single disk or set of disks.

## 2. Understanding Disk Contention

*Disk contention* occurs whenever the physical limitations of a disk drive are reached and other processes have to wait. Disk drives are mechanical and have a physical limitation on both disk seeks per second and throughput. If you exceed these limitations, you have no choice but to wait. You can find out if you are exceeding these limits both through Oracle's file I/O statistics and through operating system statistics. Although the Oracle statistics give you an accurate picture of how many I/Os have taken place for a particular data file, they may not accurately represent the entire disk because other activity outside of Oracle may be incurring disk I/Os. Remember that you must correlate the Oracle data file to the physical disk on which it resides.

Information about disk accesses is kept in the dynamic performance table V$FILESTAT.

Important information in this table is listed in the following columns:

- ***PHYRDS*:** The number of physical reads done to the data file.
- ***PHYWRTS***: The number of physical writes done to the data file.

The information in V$FILESTAT is referenced by file number. The dynamic performance table V$DATAFILE contains a reference to this number as well as other useful information such as this:

- ***NAME*:** The name of the data file.
- ***STATUS*:** The type of file and its current status.
- ***BYTES*:** The size of the data file.

Together, the V$FILESTAT and V$DATAFILE tables can give you an idea of the I/O usage of your data files. Use the following query to get this information:

```
SQL> SELECT substr(name,1,40), phyrds, phywrts, status, bytes
2 FROM v$datafile df, v$filestat fs
3 WHERE df.file# = fs.file#;
SUBSTR(NAME,1,40)                        PHYRDS PHYWRTS STATUS  BYTES
---------------------------------------- -------- -------- ------ --------
C:\UTIL\ORAWIN\DBS\wdbsys.ora             221      7      SYSTEM 10485760
C:\UTIL\ORAWIN\DBS\wdbuser.ora 0 0 ONLINE 3145728
C:\UTIL\ORAWIN\DBS\wdbrbs.ora  2  0 ONLINE 3145728
C:\UTIL\ORAWIN\DBS\wdbtemp.ora 0 0 ONLINE 2097152
```

The total I/O for each data file is the sum of the physical reads and physical writes. It is important to make sure that these I/Os don't exceed the physical limitations of any one disk. I/O throughput problems to one disk may slow down the entire system depending on what data is on that disk. It is particularly important to make sure that I/O rates are not exceeded on the disk drives.

### 3. Identifying Disk Contention Problems

To identify disk contention problems, you must analyze the I/O rates of each disk drive in the system. If you are using individual disks or disk arrays, the analysis process is slightly different. For individual disk drives, simply invoke your operating system or third-party tools and check the number of I/Os per second on an individual disk basis. This process gives you an accurate representation of the I/O rates on each drive. A general rule of thumb is not to exceed 50 I/Os per second per drive with random access, or 100 I/Os per second per drive with sequential access. If you are experiencing a disk I/O problem, you may see excessive idle CPU cycles and poor response times. For a disk array, also invoke your operating system or third-party tools and check the same items specifically the number of I/Os per second per disk. The entire disk array appears as one disk. For most popular disk arrays on the market today, it is accurate to simply divide the I/O rate by the number of disks to get the I/Os per second per disk rate. The next step in identifying a disk contention problem is to determine the I/O profile for your disk. It is sufficient to split this into two major categories: sequential and random I/O. Here is what to look for:

- **Sequential I/O.** In sequential I/O, data is written or read from the disk in order, sovery little head movement occurs. Access to the redo log files is always sequential.
- **Random I/O.** Random I/O occurs when data is accessed in different places on the disk, causing head movement. Access to data files is almost always random. For database loads, access is sequential; in most other cases (especially OLTP), the

access patterns are almost always random.

With sequential I/O, the disk can operate at a much higher rate than it can with random I/O. If *any* random I/O is being done on a disk, the disk is considered to be accessed in a random fashion. Even if you have two separate processes that access data in a sequential manner, the I/O pattern is random.

With random I/O, there is not only access to the disk but a large amount of head movement, which reduces the performance of the disks.

Finally, check these rates against the recommended I/O rates for your disk drives. Here are some good guidelines:

- **Sequential I/O.** A typical SCSI-II disk drive can support approximately 100 to 150 sequential I/Os per second.
- **Random I/O.** A typical SCSI-II disk drive can support approximately 50 to 60 random I/Os per second.

## 4. Solving Disk Contention Problems

There are a few rules of thumb you should follow in solving disk contention problems:

- **Isolate sequential I/Os.** Because sequential I/Os can occur at a much higher rate, isolating them lets you run these drives much faster.
- **Spread out random I/Os as much as possible.** You can do this by striping table data through Oracle striping, OS striping, or hardware striping.
- **Separate data and indexes.** By separating a heavily used table from its index, you allow a query to a table to access data and indexes on separate disks simultaneously.
- **Eliminate non-Oracle disk I/O from disks that contain database files.** Any other disk I/Os slow down Oracle access to these disks.

The following sections look at each of these solutions and determine how they can be accomplished.

### 4.1 Isolate Sequential I/Os

Isolating sequential I/Os allows you to drive sequentially accessed disks at a much higher rate than randomly accessed disks. Isolating sequential I/Os can be accomplished by simply putting the Oracle redo log files on separate disks. Be sure to put each redo log file on its own disk—especially the mirrored log file (if you are mirroring with Oracle). If you are mirroring with OS or hardware mirroring, the redo log files will already be on separate volumes. Although each log file is written sequentially, having the mirror on the same disk causes the disk to seek between the two log files between writes, thus degrading performance. It is important to protect redo log files against system failures by mirroring them. You can do this through Oracle itself, or by using OS or hardware fault tolerance features.

### 4.2 Spread Out Random I/Os

By the very nature of random I/Os, accesses are to vastly different places in the Oracle data files. This pattern makes it easy for random I/O problems to be alleviated by simply adding more disks to the system and spreading the Oracle tables across these disks. You can do this by striping the data across multiple drives or (depending on your configuration) by simply putting tables on different drives. *Striping* is the act of transparently dividing the contents of a large data source into smaller sources. Striping can be done through Oracle, the OS, or through hardware disk arrays.

### 4.3 Oracle Striping

Oracle striping involves dividing a table's data into small pieces and further dividing these pieces among different data files.

Oracle striping is done at the tablespace level with the CREATE TABLESPACE command. To create a striped tablespace, use a command similar to this one:

```
SQL> CREATE TABLESPACE mytablespace
2 DATAFILE 'file1.dbf' SIZE 500K,
3 'file2.dbf' SIZE 500K,
4 'file3.dbf' SIZE 500K,
5 'file4.dbf' SIZE 500K;
Tablespace created.
```

To complete this task, you must then create a table within this tablespace with four extents. This creates the table across all four data files, which (hopefully) are each on their own disk.

Create the table with a command like this one:

```
SQL> CREATE TABLE mytable
2 ( name varchar(40),
3 title varchar(20),
4 office_number number(4) )
5 TABLESPACE mytablespace
6 STORAGE ( INITIAL 495K NEXT 495K
7 MINEXTENTS 4 PCTINCREASE 0 );
Table created.
```

In this example, each data file has a size of 500K. This is called the *stripe size*. If the table is large, the stripes are also large (unless you add many stripes). Large stripes can be an advantage when you have large pieces of data within the table, such as BLOBs. In most OLTP applications, it is more advantageous to have a smaller striping factor to distribute the I/Os more evenly. The size of the data files depends on the size of your tables. Because it is difficult to manage hundreds of data files, it is not uncommon to have one data file per disk volume per tablespace. If your database is 10 gigabytes in size and you have 10 disk volumes, your data file size will be 1 gigabyte. When you add more data files of a smaller size, your I/Os are distributed more evenly, but the system is harder to manage because there are more files. you can achieve both manageability and ease of use by using a hardware or software disk array. Oracle striping can be used in conjunction with OS or hardware striping.

### 4.4 OS Striping

Depending on the operating system, striping can be done at the OS level either through an operating system facility or through a third-party application. OS striping is done at OS installation time. OS disk striping is done by taking two or more disks and creating one large *logical disk*. In sequence, the stripes appear on the first disk, then the second disk, and so on (see Figure 1).

The size of each stripe depends on the OS and the striping software you are running. To figure out which disk has the desired piece of data, the OS must keep track of where the data is. To do this, a certain amount of CPU time must be spent maintaining this information. If fault tolerance is used, even more CPU resources are required. Depending on the software you are using to stripe

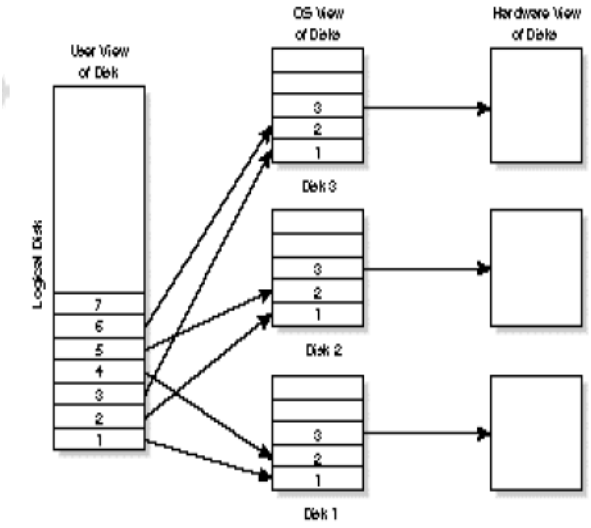


**Fig 1:** OS Striping

the disks, the OS monitoring facilities may display disk I/O rates on a per-disk basis or

on a per-logical-disk basis. Regardless of how the information is shown, you can easily determine the I/O rate per disk. Many of the OS-striping software packages on the market today can also take advantage of RAID technology to provide a measure of fault tolerance. OS striping is very good; however, It does consume system resources that hardware striping does not.

### 4.5 Hardware Striping

Hardware striping has a similar effect to OS striping. Hardware fault tolerance is obtained by replacing your disk controller with a disk array. A *disk array* is a controller that uses many disks to make one *logical* disk. The system takes a small slice of data from each of the disks in sequence to make up the larger logical disk (see Figure 2).

Hardware fault tolerance has the advantage of not taking any additional CPU or memory resources on the server. All the logic to do the striping is done at the controller level. As with OS striping, hardware striping can also take advantage of RAID technology for fault tolerance. As you can see in Figures 1 and 2, to the user and the RDBMS software, the effect is the same whether you use OS or hardware disk striping.

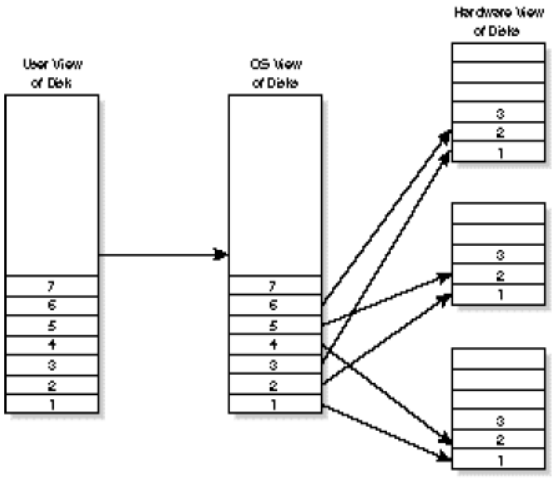


**Fig 2:** Hardware Striping

The main difference between the two is where the actual overhead of maintaining the disk array is maintained.

### 4.6 Review of Striping Options

Whether you use Oracle striping, OS striping, or hardware striping, the goal is the same: distribute the random I/Os across as many disks as possible. In this way, you can keep the number of I/Os per second requested within the bounds of the physical disks. If you use Oracle striping or OS striping, you can usually monitor the performance of each disk individually to see how hard they are being driven. If you use hardware striping, remember that the OS monitoring facilities typically see the disk volume as one logical disk. You can easily determine how hard the disks are being driven by dividing the I/O rate by the number of drives. With hardware and OS striping, the stripes are small enough that the I/Os are usually divided among the drives fairly evenly. Be sure to monitor the drives periodically to verify that you are not up against I/O limits.

Use this formula to calculate the I/O rate per drive:

| I/Os per disk = ( Number of I/Os per second per volume ) / (Number of drives in the volume) |
|---|

Suppose that you have a disk array with four drives generating 120 I/Os per second. The number of I/Os per second per disk is calculated as follows:

| I/Os per disk = 120 / 4 = 30 I/Os per second per disk |
|---|

For data volumes that are accessed randomly, you don't want to push the disks past 50 to 60 I/Os per second per disk.
To estimate how many disks you need for data volumes, use this formula:

| Number of disks = I/Os per second needed / 60 I/Os per second per disk |
|---|

If your application requires a certain data file to supply 500 I/Os per second (based on analysis and calculations), you can estimate the number of disk drives needed as follows:

```
Number of disks = 500 I/Os per second /
60 I/Os per second per disk = 16 2/3
disks or 17 disks
```

This calculation gives you a good approximation for how large to build the data volumes with no fault tolerance.

### 4.7 Separate Data and Indexes

Another way to reduce disk contention is to separate the data files from their associated indexes. Remember that disk contention is caused by multiple processes trying to obtain the same resources. For a particularly "hot" table with data that many processes try to access, the indexes associated with that data will be "hot" also.

Placing the data files and index files on different disks reduces the contention on particularly hot tables. Distributing the files also allows more concurrency by allowing simultaneous accesses to the data files and the indexes. Look at the Oracle dynamic performance tables to determine which tables and indexes are the most active.

### 4.8 Eliminate Non-Oracle Disk I/Os

Although it is not necessary to eliminate all non-Oracle I/Os, reducing significant I/Os will help performance. Most systems are tuned to handle a specific throughput requirement or response time requirement. Any additional I/Os that slow down Oracle can affect both these requirements. Another reason to reduce non-Oracle I/Os is to increase the accuracy of the Oracle dynamic performance table, V$FILESTAT. If only Oracle files are on the disks you are monitoring, the statistics in this table should be very accurate.

### 5. Reducing Unnecessary I/O Overhead

Reducing unnecessary I/O overhead can increase the throughput available for user tasks. Unnecessary overhead such as *chaining* and *migrating* of rows hurts performance. Migrating and chaining occur when an UPDATE statement increases the size of a row so that it no longer fits in the data block. When this happens, Oracle tries to find space for this new row. If a block is available with enough room, Oracle moves the entire row to that new block. This is called *migrating*. If no data block is available with enough space, Oracle splits the row into multiple pieces and stores them in several data blocks. This is called *chaining*.

### 6. Migrated and Chained Rows

Migrated rows cause overhead in the system because Oracle must spend the CPU time to find space for the row and then copy the row to the new data block. This takes both CPU time and I/Os. Therefore, any UPDATE statement that causes a migration incurs a performance penalty. Chained rows cause overhead in the system not only when they are created but each time they are accessed. A chained row requires more than one I/O to read the row. Remember that Oracle reads from the disk data blocks; each time the row is accessed, multiple blocks must be read into the SGA.

You can check for chained rows with the LIST CHAINED ROWS option of the ANALYZE command.

You can use these SQL statements to check for chained or migrated rows:

```
SQL> Rem
SQL> CREATE TABLE chained_rows (
2 owner_name varchar2(30),
3 table_name varchar2(30),
4 cluster_name varchar2(30),
5 head_rowid rowid,
6 timestamp date);
Table created.
SQL> Rem
SQL> Rem Analyze the Table in Question
```

```
SQL> Rem
SQL> ANALYZE
2 TABLE scott.emp LIST CHAINED ROWS;
Table analyzed.
SQL> Rem
SQL> Rem Check the Results
SQL> Rem
SQL> SELECT * from chained_rows;
no rows selected
```

If any rows are selected, you have either chained or migrated rows. To solve the problem of migrated rows, copy the rows in question to a temporary table, delete the rows from the initial table, and reinsert the rows into the original table from the temporary table. Run the chained-row command again to show only chained rows. If you see an abundance of chained rows, this is an indication that the Oracle database block size is too small. You may want to export the data and rebuild the database with a larger block size. You may not be able to avoid having chained rows, especially if your table has a LONG column or long CHAR or VARCHAR2 columns. If you are aware of very large columns, it can be advantageous to adjust the database block size before implementing the database. A properly sized block ensures that the blocks are used efficiently and I/Os are kept to a minimum. Don't over-build the blocks or you may end up wasting space. The block size is determined by the Oracle parameter DB_BLOCK_SIZE. Remember that the amount of memory used for database block buffers is calculated as follows:

```
Memory used = DB_BLOCK_BUFFERS (number) * DB_BLOCK_SIZE (bytes)
```

Be careful to avoid paging or swapping caused by an SGA that doesn't fit into RAM.

## 7. Dynamic Extensions

Additional I/O is generated by the extension of segments. Remember that segments are allocated for data in the database at creation time. As the table grows, extents are added to accommodate this growth. Dynamic extension not only causes additional I/Os, it also causes additional SQL statements to be executed. These additional calls, known as *recursive calls*, as well as the additional I/Os can impact performance.
You can check the number of recursive calls through the dynamic performance table, V$SYSSTAT.

Use the following command:

```
SQL> SELECT name, value
2 FROM v$SYSSTAT
3 WHERE name = 'recursive calls';
NAME VALUE
----------
recursive calls 5440
```

Check for recursive calls after your application has started running and then 15 to 20 minutes later. This information will tell you approximately how many recursive calls the application is causing. Recursive calls are also caused by the following:

- Execution of Data Definition Language statements.
- Execution of SQL statements within stored procedures, functions, packages, and anonymous PL/SQL blocks.
- Enforcement of referential integrity constraints.
- The firing of database triggers.
- Misses on the data dictionary cache.

As you can see, many other conditions can also cause recursive calls. One way to check whether you are creating extents dynamically is to check the table DBA_EXTENTS. If you see that many extents have been created, it may be time to export your data, rebuild the tablespace, and reload the data. Sizing a segment large enough to fit your data properly benefits you in two ways:

- Blocks in a single extent are contiguous and allow multiblock reads to be more effective, thus reducing I/O.
- Large extents are less likely to be dynamically extended.

Try to size your segments so that dynamic extension is generally avoided and there is adequate space for growth.

## 8. Conclusion

In this paper we have explained the impact of efficient configuration of I/O for enhancing the performance of RDBMS. For practical explanation we have used one of the popular RDBMS i.e. oracle 10g. We have suggested many parts of I/O subsystem those impact the performance of RDBMS.

## References

[1]. Lightstone, S. *et al.*, "Toward Autonomic Computing with DB2 Universal Database", *SIGMOD Record*, Vol. 31, No.3, September 2002.

[2]. Xu, X., Martin, P. and Powley, W., "Configuring Buffer Pools in DB2 UDB", IBM Canada Ltd., the National Science and Engineering Research Council (NSERC) and Communication and Information Technology Ontario (CITO), 2002.

[3]. Chaudhuri, S. (ed). Special Issue on, "Self-tuning Databases and Application Tuning", *IEEE Data Engineering, Bulletin 22(2)*, June 1999.

[4]. Bernstein, P. *et al*., "The Asilomar Report on Database Research", *ACM SIGMOD Record* 27(4), December 1998, pp. 74 - 80.

[5]. Nguyen, H. C., Ockene, A., Revell, R., and Skwish, W. J., "The role of detailed simulation in capacity planning". *IBM Syst. J.* 19, 1 (1980), 81-101.

[6]. Seaman, P. H., "Modeling considerations for predicting performance of CICS/VS systems", IBM Syst. J. 19, 1 (1980), 68-80.

[7]. Foster, D. V., Mcgehearty, P. F., Sauer, C. H., and Waggoner, C. N., "A language for analysis of queuing models", *Proceedings of the 5th Annual Pittsburgh Modeling and Simulation Conference* (Univ. of Pittsburgh, Pittsburgh, Pa., Apr. 24-26). 1974, pp. 381-386.

[8]. Reiser, M., and Sauer, C. H., "Queuing network models: Methods of solution and their program implementation", *Current Trends in Programming Methodology*. Vol. 3, Software Modeling and Its Impact on Performance, K. M. Chandy and R. T. Yeb, Eds. Prentice-Hall, Englewood Cliffs, N. J., 1978, pp. 115-167.

[9]. Borovits, I., and Neumann, S., "Computer Systems Performance Evaluation", *D.C. Heath and Co., Lexington, Mass*., 1979.

[10]. Enrique Vargas, "High Availability Fundamentals", *Sun BluePrints™ OnLine*, November 2000, http://www.sun.com/blueprints

[11]. Harry Singh, "Distributed Fault-Tolerant/High-Availability Systems", *Trillium Digital Systems, a division of Intel Corporation, 12100 Wilshire Boulevard*, Suite 1800 Los Angeles, CA 90025-7118 U.S.A. Document Number 8761019.12.

[12]. David McKinley, "High availability system. High availability system platforms", *Dedicated Systems Magazine* - 2000 Q4 (http://www.dedicated-systems.com)

[13]. Sasidhar Pendyala, "Oracle's Technologies for High Availability", Oracle Software India Ltd., India Development Centre.

[14]. James Koopmann, "Database Performance and some Christmas Cheer", an article in the *Database Journal, January* 2, 2003.

[15]. Frank Naudé, "Oracle Monitoring and Performance Tuning", http://www.orafaq.com/faqdbapf.htm

[16]. Michael Marxmeier, "Database Performance Tuning",
[17]. http://www.hpeloquence.com/support/misc/dbtuning.html
[18]. Sharma H., Shastri A., Biswas R. **"**Architecture of Automated Database Tuning Using SGA Parameters", *Database System Journal*, Romania, 2012
[19]. Sharma H., Shastri A., Biswas R. "A Framework for Automated Database TuningUsing Dynamic SGA Parameters and Basic Operating System Utilities", *Database System Journal*, Romania, 2013
[20]. Mihyar Hesson, "Database performance Issues"
[21]. PROGRESS SOFTWARE, Progress Software Professional Services,

**Hitesh Kumar Sharma:** Author is an Assistant Professor (Senior Scale) in University of Petroleum& Energy Studies, Dehradun. He has published 30+ research papers in International Journal and 10+ research papers in National Journals.

**Christalin Nelson. S:** Author is an Assistant Professor (Selection Grade) in University of Petroleum& Energy Studies, Dehradun. He has published 40+ research papers in International Journal and 12 research papers in National Journals. He is Programe Head of the computer Science Department.

**Sanjeev Kumar Singh:** Author is an Associate Professor in Galgotias University, Noida. He has published 35+ research paper in International Journal and 15+ research papers in National Journals. He is Ph.D. in Mathematics.

# A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)

Hakob GRIGORYAN
Bucharest University of Economic Studies, Bucharest, Romania
Grigoryanhakob90@yahoo.com

**Abstract:** *The research presented in this work focuses on financial time series prediction problem. The integrated prediction model based on support vector machines (SVM) with independent component analysis (ICA) (called SVM-ICA) is proposed for stock market prediction. The presented approach first uses ICA technique to extract important features from the research data, and then applies SVM technique to perform time series prediction. The results obtained from the SVM-ICA technique are compared with the results of SVM-based model without using any pre-processing step. In order to show the effectiveness of the proposed methodology, two different research data are used as illustrative examples. In experiments, the root mean square error (RMSE) measure is used to evaluate the performance of proposed models. The comparative analysis leads to the conclusion that the proposed SVM-ICA model outperforms the simple SVM-based model in forecasting task of nonstationary time series.*

**Keywords:** *support vector machines, regression, independent component analysis, financial time series, stock prediction*

## 1 Introduction

In recent years, the fast growing financial markets opened new horizons for investors and the same time bringing new challenges for financial analysts in their efforts to make effective decisions and reduce the investment risks. Stock market is a highly dynamic and complex system since there are a great number of interacting factors that affect the future prices [1]. In fact, stock market prediction means understanding which economic and non-economic factors affect market prices in order to predict the target variables based on the analysis of historical data. Thus, many researchers in the field of economic predictions have claimed that stock market prediction is a difficult task compared with other time series analysis problems as stock data is non stationary, random and chaotic [2][3]. However, financial time series are characterized by uncertainty and noise, there is some evidence that stock markets can be predictable through the use of different methods ranging from econometric to machine learning techniques [4]. Currently, advancements in different fields of applied mathematics and information technologies have led to the development of novel prediction models based on artificial intelligence techniques. Existing research indicates that statistical techniques are useful for modelling linear problems but those fail to capture the non-linear behaviour presented in financial time series as stock markets are non-linear deterministic systems [5].

In recent years, numerous machine learning-based models have been presented for time series analysis. Among them, Support Vector Machines (SVM) is a novel technique designed to solve non-linear classification and regression problems in time series analysis. SVMs are based on the structural risk minimization principle which allows them to estimate a function by minimizing an upper bound of generalization error [6]. Due to its ability to achieve a high generalization performance and testing accuracy, SVMs have been successfully applied for time series prediction domain. In regard to financial forecasting, Trafalis and Ince (2000), Tay

and Cao, (2001) introduced an application of SVM for stock market prediction, and showed promising results compared with neural network-based models. Similarly, Kim, (2003) applied SVM to forecast the stock price index and concluded that SVM can be successfully applied to stock market prediction as an alternative to neural networks. Also, Huang et al., (2005) applied SVM to forecast the movement direction of stock market, and showed that SVM has better prediction performance compared with other statistical and machine learning methods.

The main problem in stock market forecasting is the inherent noise of the financial data. By removing unwanted information from historical data we can increase forecasting accuracy and speed. Several researchers have proposed hybrid models based on SVM and feature selection techniques. For example, Cao et al, (2003) presented prediction model based on SVM technique and also three different feature extraction techniques, namely, principal component analysis (PCA), kernel PCA (KPCA) and Indepentent Component Analysis (ICA). They concluded that use of feature extraction techniques had increased prediction accuracy and among proposed techniques the best performance showed model based on KPCA and SVM techniques. Hsu et al., (2009) applied two-stage architecture for stock price prediction based on self-organizing map (SOM) and SVM. They used SOM to decompose and classify the input data and support vector machines for regression (SVR) to predict prices, and showed that their model outperforms the standard SVM-based model in stock price prediction. Lee (2009) presented a stock market trend prediction model based on support vector machine (SVM) with a hybrid feature selection method named F-score and Supported Sequential Forward Search (F_SSFS). Their experimental results showed that their proposed hybrid model outperforms neural network based model combined with other feature selection methods. Kao et al., (2013) introduced a novel combined model based on wavelet transform, multivariate adaptive regression splines (MARS), and support vector regression (SVR) to forecast stock prices, and concluded that their proposed approach outperforms other models in forecasting the stock prices.

One model can be suitable to predict a certain financial market but fail to predict another market's data as there are different factors affecting stock prices varying from one market to another. In recent years, a series of studies concerning data forecasting based on multivariate input models including different macroeconomic variables and technical indicators have been reported [15][16][17][18]. The experimental results showed that different externally determined variables based on technical and fundamental analysis are useful in prediction of stock markets data.

The main objective of this study is the investigation of the effectiveness of combined prediction model based on ICA and SVM techniques in forecasting task of noisy stock data. In such a model, first pre-processing step is used to prepare the research data and select important features by ICA method, and then SVM based prediction model is constructed based on the selected variables. The superiority of the proposed model is shown by the comparative analysis of stock market prediction model based on SVM and ICA techniques against single SVM-based prediction model without using any feature selection technique.

The remainder of this paper is organized as follows. Section 2 gives brief introduction to Support Vector Machines (SVM) and Independent Component Analysis (ICA). Section 3 describes the proposed methodology, including data collection, preparation and forecasting model. In Section 4, the experimental results together with a comparative analysis are summarized and discussed. Finally,

the concluding remarks are presented in the Section 5.

## 2 Background

### 2.1 Support Vector Machines (SVM)

Support Vector Machines (SVM) is a family of learning algorithms originated as an implementation of the structured risk minimization (SRM) principle proposed by Vapnik [19].

The basic idea behind SVM model is to represent the given examples of data as points in a high-dimensional feature space and linearly separate the feature vectors by a maximum margin hyperplane. The diagram of linearly separable SVM is depicted in figure 1. The data points closest to the maximum margin hyperplane lying on the dotted line are used to determine the regression surface. This small subsets of data points are called support vectors, while the points within the ε–insensitive zone are not important in terms of the regression function and contribute to the error loss function. In time series analysis, the application of SVMs used in regression analysis is called Support Vector Regression (SVR) [20].

Given a training data set $G = \{(x_i, y_i),\ i = 1, \ldots, l\} \subset X \times \mathbb{R}$, where $X \subset \mathbb{R}^n$ denotes the space of the input patterns, the SVR function can be expressed as:

$$f(x) = \omega \cdot \varphi(x) + b \quad (1)$$

where, $\omega \in X$ is a weight vector, $b \in \mathbb{R}$ is a bias and $\varphi$ represents the mapping function.

The objective of the SVR is to find a function $f$ that has the most $\varepsilon$ deviation from the target $y_i$ and, the most possible flat $f$. The problem can be solved by finding the small values of the Euclidian norm $\|\omega\|^2$ which can be achieved by solving the following optimization problem.

$$minimize \quad \frac{1}{2}\|\omega\|^2 \quad (2)$$

$$subject\ to \begin{cases} y_i - \omega \cdot \varphi(x_i) - b \le \varepsilon \\ \omega \cdot \varphi(x_i) + b - y_i \le \varepsilon \end{cases}$$

The optimization problem (2) is feasible when there is $f$ such that $|f(x_i) - y_i| \le \varepsilon$ for all $(x_i, y_i) \in G$. When the training data is not linearly separable, slack variables $\xi_i, \xi_i^*$ are introduced to deal with unfeasible constraints of the optimization problem (2).

The optimization problem (2) can be re-formulated as:

$$minimize \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \quad (3)$$

$$subject\ to \begin{cases} y_i - w^T x_i - b \le \varepsilon + \xi_i \\ w^T x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

where $\frac{1}{2}\|w\|^2$ is the regularization term preventing over-learning, $(\xi_i + \xi_i^*)$ is the empirical risk; and $C > 0$ is called a regularization constant which controls the trade-off between the empirical risk and regularization term.

The ε-insensitive loss function $|\xi|_\varepsilon$ can be described as:

$$|\xi|_\varepsilon = \begin{cases} 0 & if\ |\xi| < \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \quad (4)$$

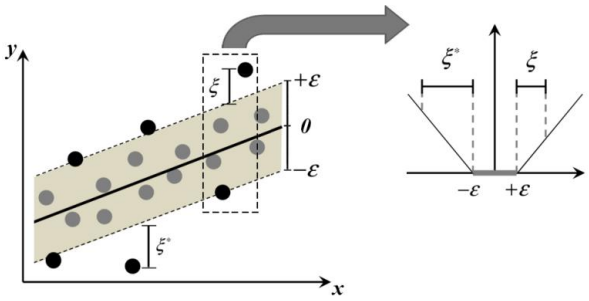


**Fig 1:** Support Vector Machines (SVM) and ε-insensitive loss function

The quadratic optimization problem (3) can be solved by introducing Lagrange multipliers.
Let *L* be the Lagrange function:

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + w^T x_i + b) - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - w^T x_i - b) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) \quad (5)$$

where $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$ are the Lagrange multipliers. Thus the dual optimization problem corresponding to (3) is given by

$$maximize \ -\frac{1}{2}\sum_{i,j=1}^{l}(a_i - \alpha_i^*)(a_j - \alpha_j^*)x_i^T x_j - \varepsilon\sum_{i=1}^{l}(a_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(a_i - \alpha_i^*) \quad (6)$$

$$subject\ to \begin{cases} \sum_{i=1}^{l}(a_i - \alpha_i^*) = 0 \\ a_i, \alpha_i^* \in [0, C] \end{cases},$$

by changing the equation $w = \sum_{i=1}^{l}(a_i - \alpha_i^*)x_i$ ,

$$f(x) = \sum_{i=1}^{l}(a_i - \alpha_i^*)x_i^T x + b \quad (7)$$

Consequently, applying Lagrange theory and the Karush-Kuhn-Tucker condition, the general SV regression function can be expressed by

$$f(x) = \sum_{i=1}^{l}(a_i - \alpha_i^*)K(x_i x) + b \quad (8)$$

where $K(x_i, x)$ is defined as kernel function. The value of kernel function is equal to the inner product of $x_i$ and $x_j$ in the feature space $\phi(x_i)$ and $(x_j)$ , such that:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (9)$$

Any function that satisfies the condition proposed by Mercer can be applied as the kernel function [19]. The most common kernel functions are Gaussian kernel and polynomial kernel functions defined as:

$$K_{gaussian} = exp(-(1/\sigma^2)(x_i - x_j)^2)$$
$$K_{polynomial} = (x_i \cdot x_j + 1)^d \quad (10)$$

where $d$ and $\sigma^2$ are the kernel parameters [21],[22].

### 2.2 Independent Component Analysis (ICA)

Independent component analysis (ICA) is an unsupervised method for extracting individual signals from a multivariate signal proposed by [23]. ICA decomposes the given dataset into components so that each component is statistically independent from the others and assumed to be non-Gaussian. The basic form of ICA model is shown in figure 2.

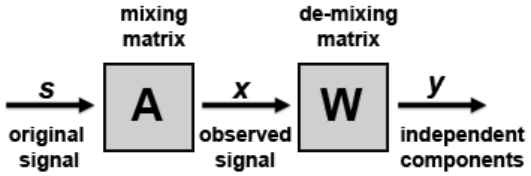


**Fig. 2.** Scheme of general ICA process

The original source signals *s* are mixed through the mixing matrix *A* to form the observed signal x, then the de-mixing matrix *W* transforms the observed signal into the independent components y.
Let $x_1, x_2, \ldots, x_n$ be the random variables, defined by the linear combinations of the random variables $s_1, s_2, \ldots, s_n$, then for any $1 \leq i \leq n$,

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n \quad (11)$$

Using vector-matrix notation and denoting the matrix A with elements $a_{i,j,}$ by $A = \|a_{i,j}\|_{1 \leq i,j \leq n}$ , and row vectors $x = [x_1, x_2, \ldots, x_m]^T$, $s = [s_1, s_2, \ldots, s_n]^T$.

The ICA model can be defined by the matrix as follows:

$$x = As = \sum_{i=1}^{n} a_i s_i \qquad (12)$$

where $a_i$ is the *i*th row of unknown matrix A, and $m \geq n$

The ICA model aims to estimate the latent variables s and unknown mixing matrix A from x with the assumption that source components $s_i$ are statistically independent and at most one of the components has a Gaussian distribution. In other words, the ICA model tends to find a de-mixing matrix W that makes the latent variables statistically independent such that,

$$y = Wx = \sum_{i=1}^{n} w_i x_i \qquad (13)$$

where $y = [y_1, y_2, \cdots, y_m]^T$ is the independent component vector, the vector $w_i$ is the *i*th row of the de-mixing matrix *W*. The elements of the vector *y* are called independent components (ICs) and they are used to estimate the source components $s_i$.

One of the ways to determine the de-mixing matrix W is the maximization of the statistical independence of ICs such as the maximization of non-Gaussianity. The non-Gaussianity of the ICs can be measured by the negentropy which is based on the concept of the information-theoretic quantity of entropy.

The negentropy of a random variable *y* with $m_y$ mean and covariance matrix $\Sigma_y$ is defined by the following equation:

$$J(y) = H(y_{gauss}) - H(y) \qquad (14)$$

where $y_{gauss}$ is a Gaussian random variable of the same covariance matrix as *y* and is distributed as $N(m_y, \Sigma_y)$ and H is the entropy of a random vector.

The entropy H is defined as:

$$H(y) = -\int p(y) \log p(y) d(y) \qquad (15)$$

where *p(y)* is the the density.

The value of negentropy is always non-negative and equal to zero when *y* has a Gaussian distribution.

The most popular algorithm to find a maximum of the non-Gaussianity of Wx is a fixed-point algorithm called FastICA [24]. The negentropy can be approximated by:

$$[E\{G(y)\} - E\{G(v)\}]^2 \qquad (16)$$

where G is any given quadratic function and $v$ is a Gaussian variable of zero mean.

By denoting g as the derivative of the nonquadratic function G, the one unit version of FastICA algorithm is as follows:

1. Choose an initial random vector $\boldsymbol{w}$,
2. Let $w^+ = E\{xg(w^t x)\} - E\{g'(w^t x)\}w$,
3. Let $w = w^+ / \|w\|$
4. If not converged, go back to 2

## 3 Research Design

### 3.1 The combined SVM-ICA model

This paper proposes a stock market forecasting model by combining SVM and ICA techniques (called SVM-ICA model). The proposed prediction framework consists of two stages.

In the first stage, the ICA technique is used to extract information from research data. ICA technique uses the observed data to convert original signals into separate independent components (ICs).

In the second stage, SVM technique is applied to forecast the stock prices. The features extracted by ICA technique are used as input variables to construct the prediction model.

In our prediction model, the future values of target variable are predicted by using the previous values of the same variable and sets of variables obtained from technical and fundamental analysis of the stock market.

The mathematical description of general prediction model is the following:

$$\hat{Y}_{(t+p)} = f\left(Y_t^{(d)}, X_t^{(d)}\right) \quad (17)$$

and

$$Y_t^{(d)} = \{Y_t, Y_{t-1}, Y_{t-2}, \dots, Y_{t-d+1}\} \quad (18)$$

$$X_t^{(d)} = \{X_t, X_{t-1}, X_{t-2}, \dots, X_{t-d+1}\} \quad (19)$$

where $\hat{Y}_{(t+p)}$ is the predicted closing price for the period p, d is the delay, $Y_t$ is the closing price at the moment of time t, and by $X_t = (X_t(1), X_t(2), \dots, X_t(n))^T$ we denote the vector entries of which are the values of the indicators having influence on $Y_t$. In our model, we consider p=1 and, as a result of correlogram analysis, d=2.

In both testing and training phases, the aim of the proposed methodology is to obtain the ICA representation of data in each input set and, based on it, to derive the proper parameters for the SVM model. In the training phase, the ICA model was applied to represent information concerning the given collection of input data. Then the SVM model is derived using ICA representation of the input data. In the testing phase the ICA model together with the SVM parameters are re-computed based on the following computation scheme.

Let $S' = \{x_1, x_2, \dots, x_T\}$ be the set of training input data,

and $S'_N = \{x_{T+1}, x_{T+2}, \dots, x_n\}$ be the set of input testing data

then by denoting $S_{T+K} = S' \cup \{x_{T+1}, x_{T+2}, \dots, x_{T+K}\}$

The algorithm used in forecasting the new, unseen yet, samples is described as follows.

Let $x_{T+K+1}$ be the new sample

For $k = 1, \dots, n - T$

Step 1. Compute the ICA model corresponding to the input data $S_{T+K}$

Step 2. Compute the SVR model to predict $\hat{Y}_{T+K}$ from (17)

Output: $\{\hat{Y}_{T+1}, \hat{Y}_{T+2} \dots \hat{Y}_n\}$

### 3.2. Data collection and preparation

To better understand the effectiveness of the proposed prediction model, two different experimental data sets are used in this paper, both based on real stock market data.

First data set is based on the historical weekly observations of a set of variables obtained from Bucharest Stock Exchange. The given data set covers the period from 3/9/2008 to 11/30/2014, a total of 350 cases of trading weeks. The research data set includes a total of 39 variables from which 35 variables were selected from technical analysis and 4 macroeconomic variables were obtained from fundamental analysis of OMV Petrom stock (symbol OMV). The closing price of OMV Petrom stock was used as a forecasting variable.

Second data set presents daily observations taken from Baltic Stock Exchange. The entire data set covers the period from 03/12/2012 to 12/30/2014, a total of 700 daily observations. The data set includes 35 variables obtained only from technical analysis of Tallink stock (symbol TALL1T). The closing price of Tallink stock was used as a target variable for prediction model.

The collected data samples have different scales as they come from different markets and sources. It is essential to consider data pre-processing by normalization in order to improve the training step and prediction results of the proposed model. Thus, the original data sets are normalized into the range of [0,1] using the formula given by

$$V = \frac{v - v_{min}}{v_{max} - v_{min}} \quad (20)$$

where $V$ is the normalized data, $v$ is the original data, $v_{max}$ *and* $v_{min}$ are maximum and minimum values of $v$.

## 4 Experimental Results and Discussion

This study integrates ICA and SVM techniques to predict the stock market closing price. The proposed SVM-ICA forecasting model was tested on two different databases.

The weekly closing price of OMV Petrom stock and daily Tallink stock closing price are used in this study as target variables. The whole data set is divided into two parts. The first part (70%) is used for training step and the second part (30%) is reserved for prediction step.

For ICA process, we consider m input time series (for the first case m=39 and for the second case m=35, based on the number of input variables that influence the closing price). Each of the time series is considered as a row formed by matrix X of size m x n (for the first case n=350 and for the second case n=700, based on the total observations of variables). The separate matrix W and the independent component matrix Y are calculated by the ICA method, where each row of Y represents an individual IC.

We adopt a fast fixed-point algorithm for ICA (FastICA) to extract independent components from collected financial time-series data. The number of ICs for two databases is as follows: 4 out of 39 variables are selected for the first data set (OMV Petrom) and 3 out of 35 variables for the second data (Tallink).

After pre-processing step, the SVM model was developed for data training. In building the prediction model, the performance of SVM depends on the accurately selected kernel function and parameters, such as, regularization constant C and loss function $\varepsilon$ defined in section 2.1. In this experiment, the Radial Basis Function (RBF) was used as a kernel function, $K(x,x') = exp(-\gamma\|x - x'\|^2), \gamma > 0$ as it is suitable for non-linear problems. In the literature of SVM, one of the suggested methods for the choice of the parameters C and $\varepsilon$ is based on the cross-validation via grid-search method proposed in [25]. The pair of parameters C and $\varepsilon$ with the best cross-validation accuracy which generate the minimum forecasting is considered the best parameter set. Then, the trained SVM model with proper parameter setting is preserved and employed in the testing phase based on the algorithm presented in Section 3.1.

For building the SVM prediction model, the LIBSVM tool box was used in this study [26].

The prediction performance is evaluated in terms of root mean squared error (RMSE), defined by:

$$RMSE(T,P) = \sqrt{\frac{1}{nr}\sum_{i-1}^{nr}\left(T(i)-P(i)\right)^2} \qquad (21)$$

where $T = \left(T(1),T(2),\dots,T(nr)\right)$ is the vector of target values, $P = \left(P(1),P(2),\dots,P(nr)\right)$ is the vector of predicted values and $nr$ is the number of data samples.

The prediction results of the proposed integrated SVM-ICA model are compared with the model based on single SVR technique without using any preprocessing step. The forecasting error is given in the table 1.

**Table 1**. The forecasting results using SVM-ICA model against single SVM model

| Data set | RMSE | Model |
|---|---|---|
| Bucharest Stock Exchange | 0.0729 | SVM |
| | 0.022593 | SVM-ICA |
| | | |
| Baltic Stock Exchange | 0.12286 | SVM |
| | 0.019191 | SVM-ICA |

The actual closing price of OMV Petrom stock and its predicted values for 105 new samples using SVM-ICA and single SVM models are depicted in figure 3 and 4. The data set was obtained from Bucharest Stock exchange with 350 weekly observations, and the prediction results are given for 105 news samples.

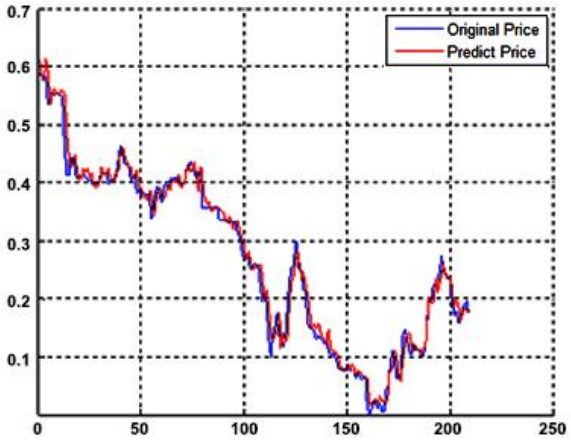


**Fig. 3.** Prediction results of OMV Petrom stock using SVM-ICA model, RMSE=0.022593.

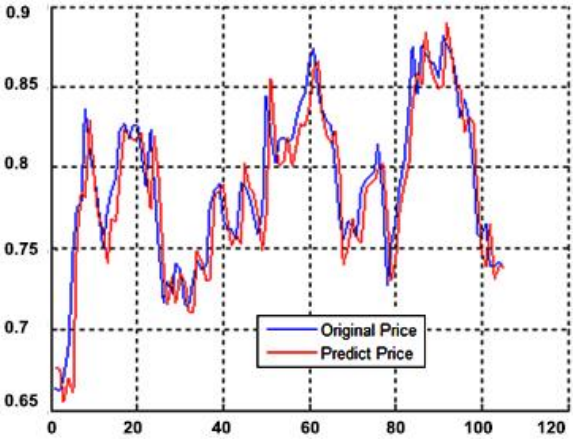


**Fig. 4.** Prediction results of OMV Petrom stock using single SVM model without any pre-processing step, RMSE=0.0729.

For the data set obtained from Baltic Stock exchange with 700 daily observations, the prediction results of the closing price of Tallink stock for 210 news samples yet not seen are depicted in the figure 5 and 6.

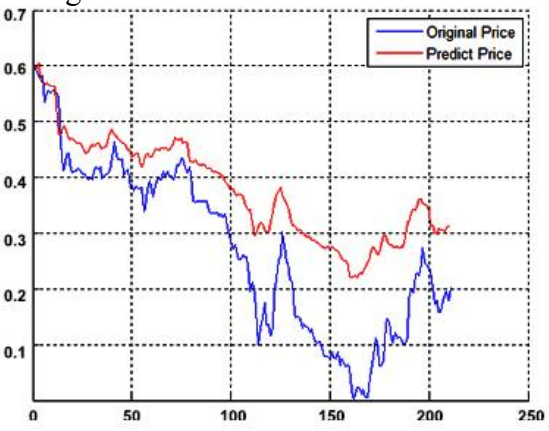


**Fig. 5.** Prediction results of Tallink stock using SVM-ICA model, RMSE=0.019191

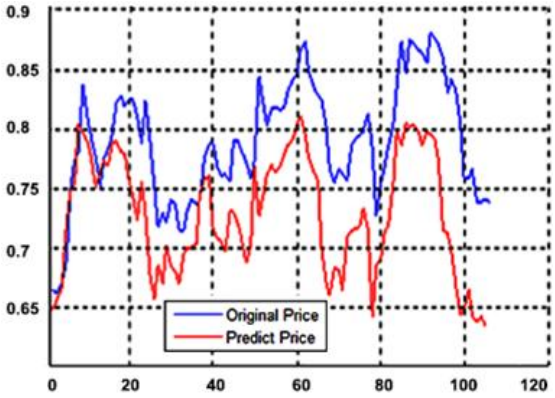


**Fig. 6.** Prediction results of Tallink stock using single SVM model without any pre-processing step, RMSE=0.12286.

## 5 Conclusion

Currently, stock market prediction is one of the challenging tasks of time-series analysis domain. In recent years, different prediction models have been presented using novel machine learning techniques. SVM is a new but promising approach for financial predictions. In this research, the combined stock market prediction model based on ICA and SVM techniques was examined. The main idea was to demonstrate the effectiveness of the hybridization of two methods in forecasting task of noisy data. This study used ICA to select input variables from technical and fundamental analysis, and SVM model to forecast the closing price. To show the effectiveness of the proposed methodology, two different data sets were used in the experiments. In addition, comparative analysis has been conducted against the model that uses only SVM technique without pre-processing step. The results obtained from the hybridized SVM-ICA model showed that ICA method effectively improved forecasting results from the point of view RMSE measure. This study allows us to conclude that SVM technique is an effective method for stock market prediction when it is combined with feature selection techniques. Moreover, experimental results obtained from proposed two-stage model encourage us to use other

pre-processing methods for gaining better experimental results.

## References


[1] Gujarati, D. N. (2003). Basic Econometrics. 4th. *New York: McGraw-Hill*.

[2] Yudong, Z., & Lenan, W. (2009). Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. Expert Systems with Applications, 36(5), 8849–8854.

[3] Abu-Mostafa, Y. S., & Atiya, A. F. (1996). Introduction to financial forecasting. *Applied Intelligence*, *6*(3), 205-213.

[4] Tsang, P. M., Kwok, P., Choy, S. O., Kwan, R., Ng, S. C., Mak, J., ... & Wong, T. L. (2007). Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence*, *20*(4), 453-461.

[5] Hill, T., O'Connor, M., & Remus, W. (1996). Neural network models for time series forecasts. *Management science*, *42*(7), 1082-1092.

[6] Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory* (Vol. 1). New York: Wiley.

[7] Trafalis, T. B., & Ince, H. (2000, July). Support vector machine for regression and applications to financial forecasting. In *ijcnn* (p. 6348). IEEE.

[8] Tay, F. E., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega*, *29*(4), 309-317.

[9] Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1), 307-319.

[10] Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, *32*(10), 2513-2522.

[11] Cao, L. J., Kok Seng Chua, W. K. Chong, H. P. Lee, and Q. M. Gu. "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. " *Neurocomputing* 55, no. 1 (2003): 321-336.

[12] Hsu, S. H., Hsieh, J. P. A., Chih, T. C., & Hsu, K. C. (2009). A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications*, *36*(4), 7947-7951.

[13] Lee, M. C. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, *36*(8), 10896-10904.

[14] Kao, L. J., Chiu, C. C., Lu, C. J., & Chang, C. H. (2013). A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting. *Decision Support Systems*, *54*(3), 1228-1244.

[15] Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert systems with Applications*, *38*(5), 5311-5319.

[16] de Oliveira, F. A., Nobre, C. N., & Zarate, L. E. (2013). Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index–Case study of PETR4, Petrobras, Brazil. *Expert Systems with Applications*, *40*(18), 7596-7606.

[17] Cocianu, C. L., & Grigoryan, H. (2015). An Artificial Neural Network for Data Forecasting Purposes. *Informatica Economica*, *19*(2), 34-45.

[18] Grigoryan, H. Stock Market Prediction using Artificial Neural Networks. Case Study of TAL1T, Nasdaq OMX Baltic Stock. *Database Systems Journal BOARD*, 14.

[19] Vapnik V.N. (2000). The Nature of Statistical Learning Theory, Springer, New York.

[20] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression.*Statistics and computing*, *14*(3), 199-222.

[21] Catalina Cocianu, Luminita State, *Kernel-Based Methods for Learning Non-Linear SVM*, Economic Computation and Economic Cybernetics Studies and Research, No. 1/2013, ISSN 0424-267X, pp. 41-60

[22] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in Proc. 2th Berkeley Symp. Mathematical Statistics and Probabilistics. Berkeley, CA: Univ. California Press, 1951, pp. 481–492.

[23] Herault, J., & Jutten, C. (1986, August). Space or time adaptive signal processing by neural network models. In *Neural networks for computing* (Vol. 151, No. 1, pp. 206-211). AIP Publishing.

[24] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Transactions on Neural Networks 10 (1999) 626–634.

[25] Hsu, C.-W., Chang, C.-C., & Lin, C.-J., (2004). A practical guide to support vector classification. Technical report. Department of Computer Science and Information Engineering, National Taiwan University.

[26] Chang, C. C., & Lin, C. J. (2001). LIBSVM: a Library for Support Vector Machines.

[29] Cocianu, C., State, Luminita, & Vlamos, P. (2009). NEURAL IMPLEMENTATION OF A CLASS OF PCA LEARNING ALGORITHMS. *Economic Computation and Economic Cybernetics Studies and Research*,*43*(3), 141-153.

**Hakob GRIGORYAN** has graduated the Faculty of Cybernetics of the State Engineering University of Armenia (Polytechnic) in 2011. In 2014 he graduated the Faculty of Informatics and Mathematics of the University of Bucharest with a specialization in Database and Web Technologies. At the present, he is earning his PhD degree in Economic Informatics at Bucharest University of Economic Studies, coordinated by Professor Catalina-Lucia Cocianu. His PhD thesis is "Machine Learning-Based Techniques for Financial Data Analysis and Forecasting Purposes".

# Forecasting mobile games' retention using Weka

Roxana Ioana STIRCU
*Bucharest University of Economic Studies*
roxana.stircu@gmail.com

**Abstract:** *In the actual market, when thousands of mobile, PC or console games are released every year, developing and publishing a successful and profitable game is a very challenging process. The gaming industry is very competitive, and all the distribution channels are full of projects competing for players. More and more companies are investing a lot of time and resources in developing an effective way to save and store all the data used and generated by their game's users. In order to develop effective and successful projects, companies adopted a lot of tools and techniques from other domains, like Statistics, Business Intelligence, or Project Management. The method most currently used is Analytics, defined as the process of discovering and communicating patterns in data, to better understand players' behavior, analyze their in-game interaction, and predicting their next in-game actions. This represents a huge step forward for the gaming industry, towards successful projects and user-tailored gaming experience. In this article the problem of users' retention is discussed, and a regression model is proposed in order to forecast players' retention, and prevent players from leaving the game.*
**Keywords**: *game analytics, metrics, user behavior, Weka, linear regression, forecast, players' retention.*

## 1 Introduction

Game Analytics was well documented and all the important concepts were defined by M. S. El-Nasr (editor) et al. [1], which covers a variety of analytics topics. The book is focused on behavioral telemetry and its role in the game development process and research. Authored by more than 50 experts in the field, the book covers data mining, visualization, monetization, and user research.

The process of user research is crucial in developing a game for success, and the first thing to mention is the frequency of the player's come backs in game, and the daily time spent playing. There is a direct connection between these two metrics and the following in-game activity of the player, as also mentioned in S. K. Hui paper [2] about gamer retention, defining retention as "a key input to gamer lifetime value".

In this paper is studied the retention and its properties, and a forecast of the next week values of retention is made. This forecast can be very useful on the long term, because developers are able to predict which users are about to quit the game, and take measures to prevent this and improve game experience.

## 2. Game analytics context

In the context of data analysis, a very important part is represented by the prediction models created using machine learning and forecasting methods, as described in [4] and [5]. Prediction represents the process of forecasting future values of a time series based on the known values, and are widely used in areas like financial markets, healthcare, marketing, social/products networks, military operations, or national economies. Prediction can be hard because of noise, or not having the right or enough data to train the model, but this problems can be resolved by using a moving average to smoothen the time series, and the data can be pre-processed and cleaned during this steps.

Machine learning represents the automatically learning process to make accurate predictions, based on previous observations (see [6] for details). The

process of classification (classify samples into predefined set of categories) can be represented as seen in *Figure 1*, listed below:


**Fig. 1.** The process of classification

Examples of classification problems can be given from a sort of areas, like natural-language processing, market segmentation, bioinformatics, face recognition, or text categorization. Among the advantages of solving problems using ML is the fact that machine learning algorithms are often more accurate than human-crafted rules, and are very flexible (can be applied to any learning task). One disadvantage is represented by the fact that is needed a lot of labeled data for the process of prediction to be more accurate.

The rules of data analysis can also be applied to data generated from video/mobile games, and this is mainly known as game analytics. The basic tools for game analysis are represented by KPI's (or Key Performance Indicators).

This are the basic metrics defined according to each game (please see [1] and [3] for detailed definitions and examples), and are the most important metrics to be tracked over time, like:

- DAU, abbreviation of Daily Active Users, defined as the total number of unique users that were active, calculated on a daily basis; active is defined as any action made in game, marked by a session – including only opening/closing the game;
- MAU, abbreviation of Monthly Active Users, defined similarly to DAU, the only difference is the fact that the total number of active users is calculated during an entire month;
- ARPU, abbreviation of Average Revenue Per User, defined as the average value of the game revenue for each user (calculated as the total revenue of the game divided by the total number of users); is also used a derivate of ARPU, named ARPPU (Average Revenue Per Paying User), calculated as the average value of the revenue only for the users that made at least one purchase in game;
- Retention, defined as the number of players that come back in the game, and it's calculated every day; most important values of retention are for the first day, the third day, and seventh day, and the thirtieth day;
- (Average) Session Length, defined as the (average) value of a player's session in game; a game session is defined by the moment when the player opens the game and the moment when he closes the game, and represents the amount of time between these two time stamps.

The most important KPI in this case is Retention, which represents the number of players that come back in game within a given timeframe. This value is crucial for any game developer, as it measures the rate of success over time for the game, according to sessions and number of players. A small value for retention indicates the fact that the players are not retained/do not come back in the game after the given period of time.

In the case of predicting a small retention value and identifying the users that are about to quit the game in the following days, some CRM campaigns or other type of player engaging actions should be taken into consideration, like:

- Promotions: represented by special offers and special prices of certain packages or boosters offered in the game's store; this can also represent an offer for a smaller price of the next month subscription or for the next mission (details are strictly defined according to the game characteristics);
- Newsletters and Invites: very important for the game features success, and is a good channel leading directly to the player's email inbox; usually via newsletter are sent articles about the game's users, development, new features, promotions, community, and invites to all the game related events;
- Rewards: this is one of the most effective ways to engage or re-engage a player in the game, and represents the process of sending rewards and boosters for free to certain users (usually this users are about to quit the game because they are having a hard time to passing a game level, and by receiving this reward they manage to advance in game).

## 3. Retention forecast

The first step in improving the game success by improving the retention rate (by campaigns) is to identify users that are about to quit the game (churn users), because can be a direct connection between decreasing user activity and churn rate.

To be able to use the information about churn users and predict retention for the following timeframe, one of the solutions can be to use machine learning based on historical tracking data. In this case we will be using historical data to train the defined model, and then apply the model on new data to obtain retention predictions and churn users.

The data used for this model should be extracted from a mobile game, using metrics defined according to each game. In this case we are interested in the following KPI:

- Sessions Per Day, defined as the total number of sessions a player has during a day, where sessions are defined as the total time between the moment when the player opens and closes the game;
- Time in game per day, defined as the total time a player spends in game during a day, and calculated as the sum of the sessions he had during that day, because these parameters represent best the retention of the players.

The data is automatically generated, using a Matlab function (*randi*). The values of sessions/day and time in game/day (in seconds), are created respecting the following conditions, displayed below in *Table 1*, where day 1 corresponds to the first day after the game installation day. In order to simulate real data generated from a game, we should use the known characteristics of users' data.

***Table 1. Average Retention values***

| Day Number | (Average) Retention Range |
|---|---|
| Day 1 | 60-65% |
| Day 3 | 50-55% |
| Day 7 | 40-45% |
| Day 30 | 20% |
| Day 60 | 5% |

We need to observe user behavior: for each user we have several time series, and the users with too few data should be ignored. The chosen metrics must explain the churn decision, in this case we defined daily numbers of sessions, and daily sessions length. The created model can be updated anytime with other metrics in order to improve performance.

For the case when the data is too noisy, we should use rolling sums over the last period of time or maybe moving averages, to eliminate the noise and made the model more accurate.

After filtering and grouping the data, we continue the process by creating the time series: the data is processed and converted to match the structure of an *.arff* file and saved with this extension, and the file is imported in Weka, to complete the preprocessing part (this is completed using the "Preprocess" tool from Weka).

To create the *.arff* file is used the structure described below, where the two attributes are defined: day, which is numeric and represents the day's number, and the time spent in game each of the days (numeric as well, counted in seconds). The data section is started with the mark "*@data*", and the attributes are listed below, separated by commas, one entry on each line of the file.

```
@relation UsersHistory

@attribute day numeric
@attribute time numeric

@data
1,4646
2,4509
3,4473
4,4357
5,4282
6,4254
7,4237
8,3767
9,3707
10,3541
...
```

We can visualize the data imported in Weka using the Visualize tool, after setting the colors, and other parameters for the selected chart. The values of daily time in game of the selected player is plotted as displayed in *Figure 2*, having the day's number displayed on OX (values from 1, representing the first day in game, to 42, representing the 42th day in game), and the total daily time spent in game, measured in seconds (values between 0 and 4865).

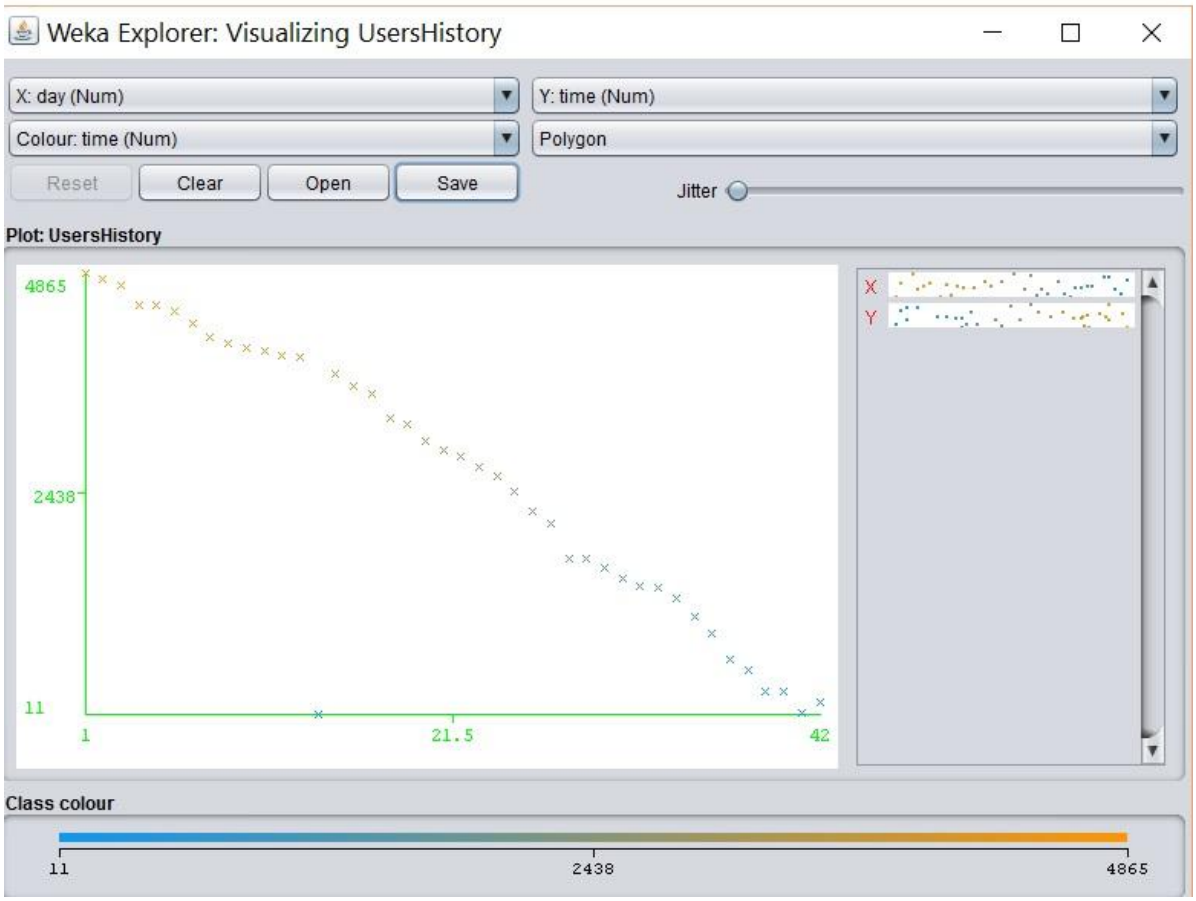


**Fig. 2**. Example of player retention during 6 weeks

For the forecast of the future values of "time in game" and retention, we use Weka's tool named "Forecast". In the first case the prediction is made using Linear

Regression as the base learner, but we also try time series forecasting with other different algorithms, for the first 6 weeks.
In order to test the forecasted values and measure the accuracy of the prediction, we compare the predicted values with the real values for the following 5 days. The forecast is made for a player with 6 weeks of history, including days with no game activity (this means that time spent in game equals 0), and the previous 4 days with no game activity at all.
The scheme used to forecast time in game is Linear Regression Model as presented in paper [4], defined below in the code section. The variable used for the regression model is the daily sessions length value, and is used to predict the expected values for the retention in the next 5 days. After choosing the desired model, *LinearRegression,* the data that it should use to build the model is loaded (the *arff* file described above), and we select "Use training set" – to specify that we want the desire model to be build based on the supplied training set. After this, we choose the dependent variable (the one we want to predict), which is the total daily time a player spends in game. The built regression model output and the predicted values are described below.

```
=== Run information ===
Scheme:
     LinearRegression -S 0 -R
1.0E-8 -num-decimal-places 4
-batch-size 1000
Lagged and derived variable
options:
     -F [time] -L 1 -M 7 -G
day -dayofweek -weekend
Relation:     UsersHistory
Instances:    42
Attributes:   2
              day
              time
Transformed training data:
              day
              time
              Lag_time-1
              Lag_time-2
              Lag_time-3
              Lag_time-4
              Lag_time-5
              Lag_time-6
              Lag_time-7
              day^2
              day^3
              day*Lag_time-1
              day*Lag_time-2
              day*Lag_time-3
              day*Lag_time-4
              day*Lag_time-5
              day*Lag_time-6
              day*Lag_time-7
time:
Linear Regression Model
time =
   -127.4628 * day +
     -0.2469 * Lag_time-5 +
      0.3209 * Lag_time-6 +
     -0.2378 * Lag_time-7 +
     -0.0099 * day*Lag_time-3
+
     -0.0107 * day*Lag_time-4
+
   6260.8175
```

The results of the forecast model described above are the predicted values for the next 5 days:

```
43*         620.361
44*          718.34
45*        321.7085
46*        115.7735
47*       -376.7138
```

Which means that the model forecasts, based on the history of the first 6 weeks, that the player may comeback in game and quit after 4 days.
We can improve this forecasting process by trying to use other learn algorithms, and compare the results afterwards for all the algorithms. Other way of improving the process can be done by speed up data processing, and by automating the prediction process.

**4. Conclusions**

This kind of model very useful to predict and prevent players churn – this thing is done in this case by selectively incentivize users that are about to quit the game. This characteristic is marked in this case by the low engagement of a player, represented in

the small values of retention vector. Using the predicted values, we can defined triggers, and automatically start campaigns and promotions, send gifts or rewards to avoid players churn, and to improve the game's success.

**References**

[1]. Magy Seif El-Nasr (Editor), Anders Drachen (Editor), Alessandro Canossa (Editor), "Game Analytics: Maximizing the Value of Player Data", Springer (2013).
[2]. Sam K. Hui, "Understanding Gamer Retention in Social Games using Aggregate DAU and MAU data: A Bayesian Data Augmentation Approach", online (2013).
[3]. Magy Seif El-Nasr, Alessandro Canossa, Anders Drachen ,"Chapter I2: Game Analytics – The Basics", online (2011).
[4]. Vasant Dhar, "Data Science and Prediction", Communications of the ACM, Vol. 56, No. 12 (2013).
[5]. John Langford, "Tutorial on Practical Prediction Theory for Classification", Journal of Machine Learning Research, 6, 273-306 (2005).
[6]. Rob Schapire, "Machine Learning Algorithms for Classification", online.

**Ioana Roxana STIRCU**, graduated the Faculty of Mathematics and Informatics of the University of Bucharest in 2010, and gained the Master title in Cryptography and Codes Theory at the same university

# A data mining approach for estimating patient demand for health services

Ionut ŢĂRANU
Bucharest University of Economic Studies
E-mail: ionut.taranu@gmail.com

**Abstract:** *The ability to better forecast demand for health services is a critical element to maintaining a stable quality of care. Knowing how certain events can impact requirements, health-care service supplier can better assign available resources to more effectively treat patients' needs.*
*The embodiment of data mining analytics can support available data to identify cyclical patterns through relevant variables, and these patterns provide actionable information to adequate decision markers at health-care structures.*
*The request for health-care services can be subject to change from time of year (seasonality) and economic factors. This paper exemplifies the efficacy of data mining analytics in identifying seasonality and economic factors as measured by time that affect patient demand for health-care services. It incorporates a neural network analytic method that is applied over a readily available dataset. The results indicate that day of week, month of year, and a yearly trend significantly impact the demand for patient services.*

**Keywords:** *Data mining, nueronal networks, decision support systems, health care IT .*

## 1 Introduction

There is increasing acknowledgment of the worth of information that exists in data resources in institution through industry sectors. Tendency, relationships between variables, and repeated models all may exist in data bases and provide accurate and deep descriptions of different processes and increase the capacity to predict and generate quantitative patterns that facilitate decision support for specialists and practitioners. A vital element to identifying patterns and relationships and generating models that facilitate simulations are multivariate techniques. A prominent field in the multivariate arena imply data mining methods that include mathematical functions and algorithms that process data resources in order to extract actionable facts for decision makers. This process of information extraction through data mining is often referred to as knowledge discovery [1] or, in other words, the recognition of valuable information that enhances facts, information, and skills for those who make decisions. The notion of leveraging data resources with mining methods to augment decision making via know-how discovery is becoming an acute component of organizational capability bearing in mind the evolving era of big and new data resources. Data resources are increasing every year in the light of the introduction of new technologies across industry sectors.

The health-care industry is confronting a meaningful growth in data resources due to the continuous progress of the digital age. The creation of electronic medical records, the process of e-prescribing, medical devices that automatically download patient elements, and the increased usage of information systems at the private practitioner and hospital and health systems level are easing the initiation of vast resources that can supply information to increase yield in a number of applications. Data mining analytics are being applied in the health-care industry across a mixture of areas. Some of these include the analysis of workflow operations of large health-care provider companies that include studies that examine the drivers of patient duration of stay, patient request and bottlenecks in

emergency room throughput, and patient satisfaction rates. Other sectors involve risk bedding applications or the better identification of patient populations at danger of developing chronic disease. Ultimately, semantic mining applications are being applied to electronic health records to better comprehend treatment and outcomes and patient diagnosis. One particular data mining technique involves the use of neural networks, which is an approach that include algorithms that processes historical data to detect both linear and non-linear patterns. The resulting models can then be used to lead 'what if' simulations on out of sample data, new data, and prognosis data. Neural networks have been used in a variety of industry applications that include the prediction of bank failures, traffic patterns, and even precipitations [2][3]. Neural networks and multivariate techniques are also being incorporated in the health-care industry to aid in knowledge discovery in a number of applications that comprise treatment efficaciousness [4] and general operations of health-care agencies [5] and patient throughput in emergency rooms.

## 2. Advanced analytics and health-care services

A significant factor in obtaining increased performance in providing services in any industry is the capability to better perceive what conduct demand for these services. With this information, decision makers can more specifically enforce the best or most favorable number of resources that are necessary to satisfy different quantity of request. Predictive analytic methods can improve the accuracy of estimating patient demand for organizations that enhance health-care services. Through analyzing patient services data in the readily available database, it was determined that seasonality factors together with general macroeconomic course had noteworthy effects on the demand for health-care services. With this information, decision makers can more precisely apply existing resources to satisfy patient expectations, and better handle costs while providing a more consistent service. In the quest to increase efficiencies, companies detect a process or functional zone that can be improved with respect to resource allocations that execute some type of task. Decision makers study the situation at hand and regulate employee pools, technological infrastructure, and compatible operations.

### 2.1. Data mining and predictive modeling in patient centered decision support

Health-care informatics techniques are crucial in awareness and supporting health-care delivery components. Data mining and predictive modeling techniques are fundamental to this because of significant improvements in information technology as well as data collection and aggregation of disparate data sources.

Medical diagnostic decision support (MDDS) systems have been used for decades. These systems have been developed because it is well-known that health-care providers are frequently asked to make crucial clinical assessment based on imprecise and/or deficient patient information. Inadequate information leads to faults, which can dramatically influence quality of care. For example, treatment for patients suffering from diabetes is exacerbated by the presence of comorbid conditions, social support challenges, and poor medication adherence. . This type of electronic decision support system ensures the appropriate implementation of evidence-based chronic care models.

Data mining methods have been used to identify the socio-demographic, physical, and psychological factors most important to the early detection and treatment of serious health-care conditions. Penny & Smith [6] explored data mining techniques to improve the quality of life of patients suffering from irritable bowel syndrome (IBS). This longitudinal cohort study examined logistic regression,

classification, and neural network models. These models demonstrated that IBS severity, psychological morbidity, marital status, and employment status significantly influenced a patient's health-related quality of life. These results provide the best information to afford better assessment and management of patients with IBS.

Other studies have examined data mining methods to improve the accuracy of diagnostic systems based on information derived from multiple, disparate data sources as well as recognition of the uniqueness of health-care data mining methods and techniques [7]. In addition, with advances in information technology, it is now possible to combine data from electronic medical records with human knowledge (i.e., expert information) to optimize the accuracy of diagnostic systems. Prediction of the onset of liver cancer [8], classification of malignant colorectal tumors and abnormal livers [9], and prediction of mortality of patients with cardiovascular disease [10] are now commonplace.

Other important applications of health-care data mining and predictive modeling techniques are resource allocation and request management in the emergency department and hospital setting. Sun et al [11] developed forecasting models to determine the probability of a hospital admission based on information collected at the point of emergency department triage. Examining 2 years of hospital data collected by nurses from emergency department patients at the point of triage, regression models were developed to determine the strongest factors in precisely predicting a patient's immediate inpatient admission from the emergency department. Outside of the obvious admission criteria (e.g., heart attack, life-threatening trauma), it is not always clear that a patient will be admitted at the point of emergency department triage for conditions such as respiratory infections, pleurisy, or orthopedic concerns. The results from this study demonstrated that age, patient acuity category, and emergency department arrival mode were the strongest predictors for hospitalization. These predictive models, if used at the point of triage, could be used for early admission planning and resource challenges faced by inpatient and acute care facilities. Similarly, there have been several studies that demonstrate the effectiveness of data mining techniques in forecasting hospital admissions, returns to the emergency department, demand for specific illnesses, same-day admissions, and emergency department demand [12],[13],[14],[15].

### 2.2 Seasonality and estimating the demand

The notion of identifying repetitive or cyclical direction in time for demand of particular processes is frequently referred to as identifying seasonal patterns of demand.

Traditionally, companies apply analytics to establish two main sources of information concerning seasonality: whether demand for their products or services has seasonal patterns (e.g. do their sales increase or decrease according to a particular point in time on a repetitive basis) and, if seasonality exists, what is the size of the change in demand according to a particular point in time. Prior research has shown that seasonality effects for patient demand for healthcare exist. For example, the 'winter effect' has been cited as being associated with increases in depression-related ailments [15],[16]. Studies have also concluded that other factors such as general economic distress (e.g. unemployment, financial stress) drive demand for mental health services [18],[19].

Analytic techniques have been used to model the impact of seasonality on patient request for health services in order to better predict future demand and allocate resources consequently. Existing research incorporating calendar-based data has concluded that seasonality provides valuable decision support for the

estimation of patient demand for urgent care clinics and emergency room facilities. Step-wise linear regression was applied to daily patient volume, which was matched with calendar data (e.g. day of week and month of year) and weather data to forecast the number of patients searching urgent care [5]. The results indicated that regression models incorporating calendar data were useful in estimating future patient demand while weather data only provided peripheral improvement to the analysis. Time series methods, linear regression, and neural network methods incorporating daily patient visits and calendar data have been utilized to study patient visits to emergency room facilities [20],[21]. This application is seen as particularly useful in helping alleviate overcrowding and enhance staffing and patient throughput by providing predictive information of patient demand [22].

### 2.3. Data and analytic methodology

In this chapter, we summarize the data collected from outpatient (patient who receives medical treatment without being admitted to a hospital) clinics that included ENT, dermatology, paediatrics, orthopaedics, and OBGYN clinics.

Table 1 classify the monthly and intra-week seasonality indices. These indices are calculated by the ratio of recurrent demand to mean demand. The coefficient of variation ($Cv= \sigma / \mu$) is calculated as a relative measure on the levels of seasonality through the clinics. As one would expect, differences can be found in both the patterns and the levels of monthly seasonality for various types of specialties. For example, request is increased during the summer time for dermatology, whilst the opposite is true for ENT. Then again, from the the intra-week models presented in Table 2 we can see that are less likely to be affected by the type of specialty. A general pattern reveals peaks on Mondays, followed by Thursdays next. We can usually notice lower demands Tuesdays and Wednesdays. Intra-day variations existed with peaks around mid-mornings and mid-afternoons.

**Table 2.1** .Seasonality data - Monthly seasonal indices

| Monthly seasonal indices | ENT | Orthopaedics | Paediatrics | Dermatology | OBGYN |
|---|---|---|---|---|---|
| January | 1.557 | 1.326 | 1.241 | 0.919 | 1.250 |
| February | 1.413 | 1.285 | 1.089 | 1.027 | 1.293 |
| March | 1.654 | 1.123 | 1.093 | 1.459 | 1.116 |
| April | 1.307 | 1.164 | 1.048 | 1.363 | 1.178 |
| May | 0.953 | 1.367 | 1.108 | 1.303 | 1.256 |
| June | 0.999 | 1.184 | 1.205 | 1.952 | 1.122 |
| July | 0.840 | 1.245 | 1.044 | 1.134 | 0.916 |
| August | 0.908 | 1.529 | 1.037 | 1.183 | 1.328 |
| September | 1.052 | 1.266 | 1.395 | 1.496 | 1.122 |
| October | 1.307 | 1.225 | 1.486 | 1.063 | 1.384 |
| November | 1.365 | 1.034 | 1.422 | 0.866 | 1.428 |
| December | 1.403 | 1.014 | 1.593 | 0.998 | 1.367 |
| *Coefficient of variation (Cv)* | 0.269 | 0.143 | 0.196 | 0.306 | 0.145 |

**Table 2.2** Seasonality data - Intra-week seasonal indices

| Intra-week seasonal indices | ENT | Orthopaedics | Paediatrics | Dermatology | OBGYN |
|---|---|---|---|---|---|
| Monday | 1.498 | 1.400 | 1.455 | 1.431 | 1.380 |
| Tuesday | 1.130 | 1.239 | 1.054 | 1.270 | 1.220 |
| Wednesday | 1.170 | 1.062 | 1.110 | 0.891 | 0.988 |
| Thursday | 1.225 | 1.215 | 1.211 | 1.446 | 1.281 |
| Friday | 1.124 | 1.231 | 1.319 | 1.076 | 1.280 |
| *Coefficient of variation (Cv)* | 0.126 | 0.119 | 0.161 | 0.226 | 0.146 |

**Neural network analysis and results**

The neural network methodology in this case allude to the utilization of complex computer algorithms that detect existing models and relationships within historical data. The neural network framework used in this analysis incorporates a multilayered perceptron[23]with a feedforward backpropagation testing function [24] .

The neural network modeling process begins with an input layer that includes nodes that correspond to each independent (driver) variable.

Driver variables are assigned weights by the algorithm, where the weighted sum of these inputs is passed into a squashing function in the hidden layer where non-linear calculations are performed on the variables relative to the dependent variable. The combined results in the input and hidden layers are passed to an output layer and compared with the historical dependent variable. Weights for variables are estimated by the backpropagation training method.

The final model is a set of code that involves a allowance or adjustment made in order to take account of special circumstances or compensate for a distorting factor scheme for independent/driver variables. Neural networks can be compared with regression analysis, with a major differentiator being that the n-net approach is based in algorithmic processing that incorporates a dynamic weighting mechanism.

**3. Conclusions**

These advanced analytic methods go across simple recognition of retrospective capabilities of prime reporting and supply decision markers with quantitative models that describe relationships among variables underpinning processes. These models provide simulation capabilities to project eventual outcomes given customization to process sariable inputs.

More unpretentiously put, analytic methods such as neural networks process historical data and determine whether there are reliable steadiness in the the rate at which things occurs or is repeated over a particular period of time and magnitude of occurrences in that data [25]. In this case do Mondays or Tuesdays of every week or particular months over the 3 years entail significant trends/patterns regarding demand for patient services.

The yield for the multivariate approach could return practical value information for hospital staffing office. The output could spot whether a particular day of the week consistently experiences +/- average demand, and would also provide an assessment of the detailed level of the demand. With this knowledge, health-care staffing operations can better keep suitable clinicians on a daily basis with greater precision to facilitate solid care for patients. In the case at hand, significant seasonal patterns were identified.

**References**

**[1]** Fayyad U, Shapiro G and Smyth P (1996) From data mining to knowledge discovery in databases. Artificial Intelligence 13(3), 37–54.

[2] Tam K and Kiang M (1992) Managerial applications of neural networks: the case of bank failure predictions. Management Science 38(7), 926–947. | Article |

[3] French M, Krajewski R and Cuykendall R (1992) Rainfall forecasting in space and time using a neural network. Journal of Hydrology 137(1–4), 1–31. | Article

[4] Lisboa P et al (2008) Time-to-event analysis with artificial neural networks: an integrated analytical and rule-based study for breast cancer. Neural Networks 21(2–3), 414–426. | Article | PubMed |

[5] Batal H, Tench J, McMillan S, Adams J and Mehler PS (2001) Predicting patient visits to an urgent care clinic using calendar variables. Academic Emergency Medicine 8(1), 48–53. | Article | PubMed |

[6] Penny KI and Smith GD (2012) The use of data-mining to identify indicators of health-related quality of life in patients with irritable bowel syndrome. Journal of Clinical Nursing 21: 2761–2771. | Article | PubMed |

[7] Cios KJ and More GW (2002) Uniqueness of medical data mining. Artificial Intelligence in Medicine 26(1–2), 1–24. | Article | PubMed |

[8] Kuo MH, Hung CM, Barnett J and Pinheiro F (2012) Assessing the feasibility of data mining techniques for early liver cancer detection. Studies in Health Technology and Informatics 180: 584–588. | PubMed

[9] Gao P et al (2012) Which is a more accurate predictor in colorectal survival analysis? Nine data mining algorithms vs. the TNM staging system. PloS One 7(7), e42015. Epub 25 July. | Article | PubMed |

[10] Austin PC, Lee DS, Steyerberg EW and Tu JV (2012) Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? Biometrical Journal 54: 657–673. | Article | PubMed |

[11] Sun Y, Heng BH, Tay SY and Seow E (2011) Predicting hospital admissions at emergency department triage using routine administrative data. Academic Emergency Medicine 18(8), 844–850. | Article | PubMed |

[12] Peck JS, Benneyan JC, Nightingale DJ and Gaehde SA (2012) Predicting emergency department inpatient admissions to improve same-day patient flow. Academic Emergency Medicine 19(9), E1045–E1054. | Article | PubMed |

[13] LaMantia MA et al (2010) Predicting hospital admission and returns to the emergency department for elderly patients. Academic Emergency Medicine 17(3), 252–259. | Article | PubMed |

[14] Kudyba S (2012) Utilizing multivariate analytics for decision support to enhance patient demand forecasting. International Institute for Analytics October, 1–7.

[15] Jones SS et al (2009) A multivariate time series approach to modeling and forecasting demand in the emergency department. Journal of Biomedical Informatics 42(1), 123–139. | Article | PubMed |

[16] Lurie SJ, Gawinski B, Pierce D and Rousseau SJ (2006) Seasonal affective disorder. American Family Physician 74(9), 1521–1524. | PubMed |

[17] Fullerton KJ and Crawford V (1999) The winter bed crisis – quantifying seasonal affects on hospital bed usage. QJM 92(4), 199–206. | Article | PubMed |

[18] Catalano R (1991) The health effects of economic insecurity. American Journal of Public Health 18(9), 1448–1152.

[19] Dooley D, Prause J and Ham-Rowbottom K (2000) Underemployment and depression: longitudinal relationships. Journal of Health and Social Behavior 41(4), 421–436. | Article | PubMed |

[20] Holleman DR, Bowling RL and Gathy C (1996) Predicting daily visits to a walk-in clinic and emergency department using calendar and weather data. Journal of General Internal Medicine 15(4), 237–239. | Article |

[21] Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ and Snow GL (2008) Forecasting daily patient volumes in the emergency department. Academic Emergency Medicine 15(2), 159–170. | Article | PubMed |

[22] Wargon M, Guidet B, Hoang TD and Hejblum G (2009) A systematic review of models for forecasting the number of emergency department visits. Emergency Medicine Journal 26(6), 395–399. | Article | PubMed |

[23] Rumelhart D, Hinton G and Williams RJ (1986) Learning internal representations by error propagation. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition. (Rumelhart DE, McClelland JL and Corporate PDP Research Group, Eds) MIT Press.
[24] Hinton G (1992) How Neural Networks Learn from Experience. Scientific American 267(3), 144–151. | Article | PubMed |
[25] Chu C and Zhang G (2003) A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of Production Economics 86(3), 217–231. | Article |

**Ionuţ Ţăranu** graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 1996, having its Master degree on "Database support for business". At present is in the process of getting his title of doctor in economy in the specialty of "Soft-computing methods for early medical diagnosis". He has been an Assistant Professor for 4 years at "Titu Maiorescu" University and also for 4 years at Academy of Economic Studies from Bucharest. He published a series of articles, from which the most important are Applying ABCD Rule of Dermatoscopy using cognitive systems and ABCDE Rule in Dermoscopy – Registration and determining the impact of parameter E for evolution in diagnosing skin cancer using soft computing alghorithms.
Mr. Ţăranu is currently the General Manager of Stima Soft company. He has more than 15 years of experience as a project manager and a business analyst with over 13 years of expertise in Software development, Business Process Management, Enterprise Architecture design and Outsourcing services. He is also involved in research projects, from which the most relevant are:

- Development of an Intelligent System for predicting, analyzing and monitoring performance indicators of technological and business processes in renewable energy area;
- Development of an eHealth platform for improving quality of life and the personalization of therapy at patients with diabetes;
- Development of an Educational Portal and integrated electronic system of education at the University of Medicine and Pharmacy "Carol Davila" to develop medical performance in dermatological oncology field;

# Implementing Business Intelligence System - Case Study

Yasser AL-HADAD[1], Răzvan Daniel ZOTA[2]
Faculty of Cybernetics, Statistics and Economic Informatics, Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, Romania
[1]dukeyasser10@yahoo.com, [2]zota@ase.ro

**Abstract:** *Understanding and analysis data is essential for making decision within a system. Any analytical tasks can be implemented directly by the transactional system but it becomes more difficult as the transactional system grows. Analytical systems and their extension appear as a solution for complex and large datasets. We think that it's time for medium companies to get the benefit from such systems as analytical systems become more variant and in hand for every possible user. In this paper, we propose an architecture of analytical system that can adapt and integrate with existent transactional system of timber export company. The proposed analytical system should have the ability of implementing the tasks required by the decision makers of the system. Also, we try to explore the ability of SQL server of implementing our proposed architecture.*

**Keywords:** *BI(Business intelligence) , DSS(*Decision support systems*), DW(Data warehouse), OLAP(*Online Analytical Processing*), ETL(*Extract, transform and load*), SAS(SQL server services)*

## 1 Introduction

Business Intelligence (BI) is high business application tools [1] used for collecting, cleansing, processing and analyzing data. evaluating and understanding the results is essential for the efficiency of decisional system. Created knowledge are collected from both internal and external sources [2]. and accumulation them can lead to improve corporate profitability [3]. BI tools becomes more variant and easy to use [4] and their solutions are ranked as one of most important technological stuffs by chief information officers [5]. Investments in BI focuses on achieving business targeted and increase return on investments [6]. So BI is an entire concept that is used for implementing a decision support system (DSS). DSS can be described as the next generation systems that follow transactional and operational systems [7]. Collected data and maybe the resulted information are stored in data warehouse (DW). It is one of the basic component of DSS and BI systems. It includes all spread data over the organization in single pool [8]. So the existing of DW was necessary as the transactional databases were unable to store the accumulated historical data [9]. DW can be accessed by proper data-analyzing to analyze and store required data [10,11]. DW help to achieve strategic business objectives by offering clean and homogeneous data in real time to support the analytical processes [12,13]. there are many architecture that are suit the design and implementation of DW [10,14,15]. The common idea is to load the propagated data from operational DBMS (database management system) and other sources in DW using a special tools called ETL (extract, transform and load). ETL is the most usable tool for integrating data into data warehouse. Many architectures are proposed for implementing ETL. It was used to integrate the entire data source in [16] even if it is too big. In [17], the data are processed in local repositories before integrating them in the global data warehouse. The difference between ETL and ELT (extract, load and transform) is presented at [18]. In [19], an arhitecture is proposed for integrating and validation data from multiple data sources. Other approach is presented in [20] that improves the

performance and interacts more with users. At [21], the metadata of ETL is discussed to enable the interoperability between the systems. The incomplete data are treated as grey data and special procedures are proposed to deal with them [22]. Data integration includes data from a single or multiple sources. The problems of data integration from multiple sources are overlapping data and matching records belongs to same entity. The challenges of data integration from single source using backup data files is similar to data integration from multiple sources. Combining integration step with cleansing step can increase the performance. on the other hand, these transformation lose the legacy data form. Legacy data is important for reevaluation of data cleansing process or backflow of cleaned data.

Timely decision making becomes difficult due to the inefficiency of transactional databases to handle the amount of information access, retrieval, update and maintenance. This shortness impact every industry [12]. OLAP (Online Analytical Processing) appears as one of solutions for resolving problems of operational database. Upon the OLAP, many research is built on data cube operator [23]. Also various algorithms for supporting dimension hierarchies [24]. OLAP is a technology that offers users to perform analysis on detailed data from multi prospects. It adopts multidimensional approach for implementing the analytical database. Analytical database analyze complex relation between millions of records for identifying trends and patterns. It stores static data type such as derived and calculated data for providing them in real time whenever there are requested by managers. The perspective of data cube and multidimensional database was adopted by many corporations such as Microsoft [25]. Now, SSAS(SQL Server Analysis Services) is the leader for business logic analysis in services and software. Its integrated platform is one of largest framework that implement the concept of BI [8]. SAS solution provides full end-to-end technology. Also for data quality and data access, it ensures that large volumes of data are processed and transformed into accurate analysis information and reporting. One of the advantages of SAS for many users are reporting and graphical representations. It offers dynamic views of information, that can be displayed in many forms such as a table, a report and a chart. The main requirement of potential users of Business intelligence solution is to perform analysis of data and testing the results. SAS Enterprise BI Server provides many tools for analysis and forecasting, which is essential to solving problems in order to make the company more competitive. One of the advantages of SAS Enterprise BI Server is the features of web portal which provides some functionalities that enable users to manipulate the contents and layout of the interface. Also it is offers features Web-based distribution and reporting. Users can manage and build reports using custom and exporting them to EXCEL or PDF files. There are functionalities that allows users to observe the results and patterns that cannot be observed in a traditional chart using more interactive methods such as 3D videos and tables with bubbles containing data presentations. SAS product offers an integrated platform that store data and reports. Also its predefined functions allow the developers of IT departments to focus on other tasks such as security of BI, implementation and maintenance.

The objective of the application is to build a decision support system (DSS) for inventory management. The system will be implemented for a medium level company that exporting timber. The objective of the project is to find an appropriate solutions for the firm size and to be adaptable for the business field and decision maker requirements. It is expected that the proposed system will provide timely relevant information to meet the needs of

decision makers so they can respond quickly to market fluctuations through readapt policies and strategies.

## 2 Development methodologies

Project development is a process that consisting of several different stages. Each stage has its own requirements and targets. Depending on the project type, certain stages gain additional attention in the overall effort. There are many methodologies types that can be used for project development. Software development methodologies can be defined as set of guidelines and rules that are used in the process of each stage. Each of these methodologies has its Characteristics, strengths and weaknesses [26]. For our case, we have used Prototyping methodology, see figure 1. Prototyping is a methodology that entails building a prototype or demo version of the software product that contains the critical functionality. Demo version includes sufficient information to build a prototype and should be built fast. It is used to refine specifications and it acts as baseline for entire project. it makes the communication process better between project owner and project team [27]. The main characteristic of prototype methodology is that project owner and users is actively involved by evaluating prototypes that are valued over writing specifications and meant to be discarded later.

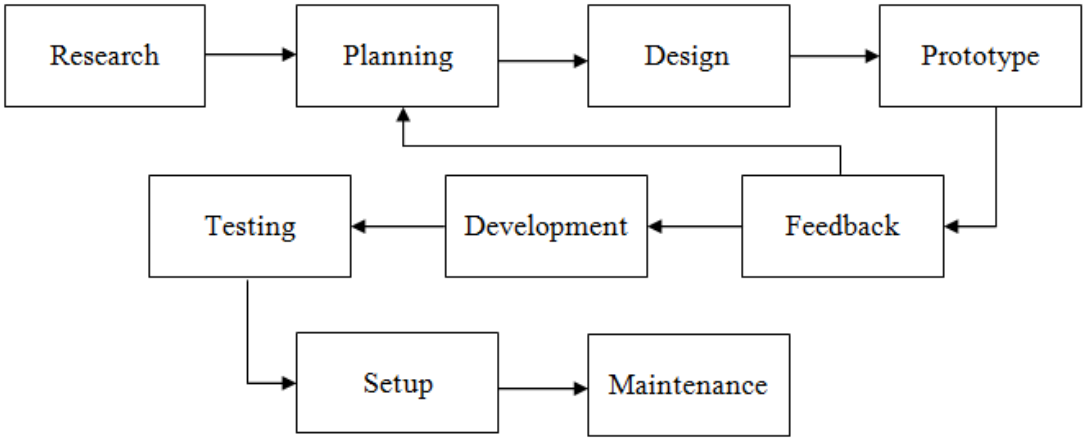


**Fig. 1.** Prototyping methodology

Demo versions can contain one or more prototypes of the initial models of the software product. In the other hand, Strengths points of this methodology is crucial for good implementation of software project. It involved the potential users of the system in process of implementation and improve their experience. Also the early feedback from users and the project owner help to early identification of any redundant or missing so an accurate identification of software requirements is guaranteed. The main Weaknesses of the prototype methodology is the costs generated by increased programming effort due to building the prototype, but this Weaknesses can be ignored because SAS platform that is used for implementing our project guaranteed that the programming effort by using SAS predefined functions and layouts will be reduced so the process of building any prototype will be easier.

## 3 Staging method

Staging method removes the performance overhead involved in cross-database joins by creating staging tables in the warehouse database. Staging method extracts data from the sources and loaded them into data warehouse without any transformation. It just makes another copy of legacy data content and loading them into analytical system in order to make them available for

any transformation processes that can be performed later. Obvious method combines both extract and transform steps in the same flow. It is more efficient and no need to store intermediate results. For an mid-level enterprise, the entire size of data is not too large to generate any performance difficulties, so the processing efficiency of the system is not the most important factor of the system. In our case, the cost of developing and maintenance is much important. as it is seen in figure 2, the flow of staging method is divided in many small steps in order to ease and simply the process of analyzing and implementation. Using obvious method, data are processed directly and loaded in the analytical database or OLAP.

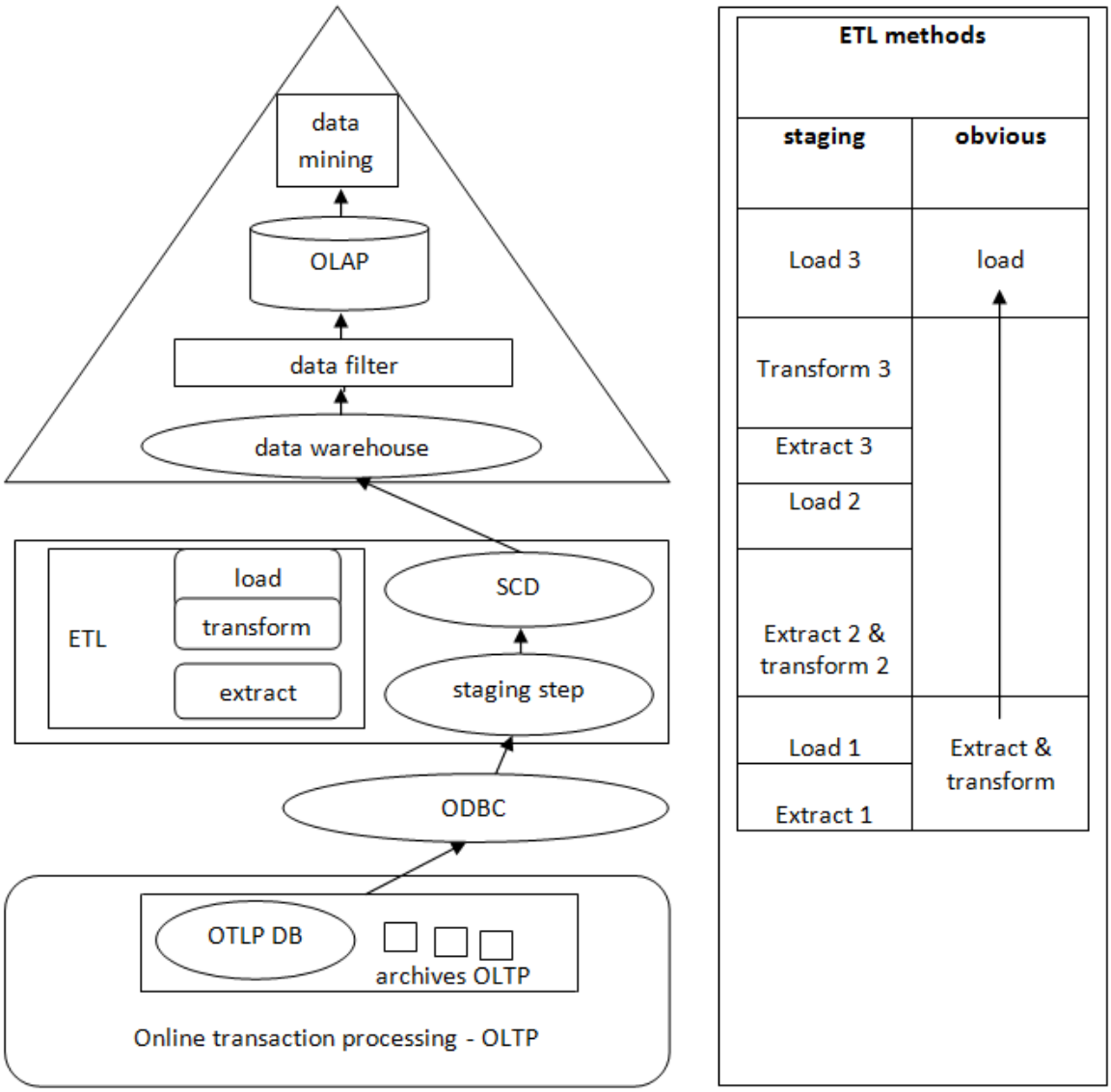


**Fig. 2.** DSS framework components

## 4 Case study

In the following section, we are going to propose DSS for our case study. The requirement of system will be analyze according to standpoint of intermediate dealer (quality, production capacity, price, payment conditions , etc). So there is no need to analyze the process of timber production as the modelled system can validate itself by using data in the transactional database where all the specific and phenomenon of the this field business are hidden in the data. A database for an timber export company with annual turnover

of over two million euro will be used for validation the proposed DSS.

### 4.1 Exiting transactional system

As the volume of data and activities increase, more challenges appear with respect to process and manipulate collected data. Existing informational system is designed to meet the requirement of transferring big part of processing data burden to an centralized system for operating the firm activities. It replaces the older system which were compose by many uncorrelated systems and procedures. That limitation was preventing the organization to expand their activity. Existing system has proposed to meet the requirement of company. Standardized solution is adopted for such modern system, but other special solutions that adapt some specific requirements.

### 4.2 Technical equipments

Transactional system is implemented using PHP scripting language, web interface is running by apache server and MySQL server for database. XAMPP server is used to run the application and database. The version XAMPP 1.6.3a is kept up to now to guarantee the working of all facilities of the system. XAMPP server database and application are installed at every unit to enable operating the system at offline mode so the server is installed at every unit. It should be configured properly at every unit to assure that is use an un busy socket and an existing hard disk partition and the platform has the full access to use the partition. In order to operate the system at the offline mode, users has to synchronize their local database using defined procedure for updating database. Synchronizing database is performing via archived files that include the latest data. For supporting the partners of corporation and agents who don't have access to informational system, system offers the possibility of exporting its data using portable document format (PDF).

### 4.3 Instruments of proposed system

SSAS(SQL Server Analysis service) is adopted for implementation the proposed analysis application. It offers the main methods and algorithms for analysis purposes. The friendly user interface facilitate the use of its services. Also the wide range of graphical representation for data and information that are available in their services makes viewing and understanding the resulted data more easier. The step of building the model should be followed by data injection step. All the resource data are available in the MySQL server that are not accessible directly for an analysis model. Any input data should be locating in the engine database SQL server. MyODBC driver connecter is used to make the data in MYSQL server accessible for integrating them in SQL server. ODBC is a standardized API (application-programming interface) that enable the client-side application to connect to one or multiple databases. MyODBC driver is a member of MySQL ODBC. It provide access to MySQL database throw the standard ODBC. It offers standard interface using driver-manager based and a native interfaces. For Unix and Mac operation system, native MySQL network can be used to communicate with MySQL database. Installing of MySQL connector ODBC is necessary for Windows and application that use the ODBC interface. Figure 3 shows the instruments are used for implementing the application as following:

- *MySQL Server*: represents the platform that hosts the current transactional database. It also used to reload the archived data into database in order to send them to SQL server platform;
- *MyODBC connecter*: offers the possibility of accessing MySQL databases directly using SSIS (SQL server integrating services). Data and its structure can be viewed and imported in the SQL server. So there is

no need to export any standard format to be imported in destination;

- *SSIS (SQL server integration services)*: offers a wide range of components that can be used for interrogating and integrating the data. The graphical user interface makes it easy to build and configure an integrating services project. This types of projects can deal with many external data resources;
- *SQL server database engine*: The data destination of any integrating services project is in SQL database engine. It is important to load all necessary data of analyzing process to make it accessible to OLAP process and analysis services;
- *SSAS (SQL server analysis services)*: design prediction or classification models for the entities available in the database for supporting decision system or any simulation system. Furthermore, OLAP process can contribute in implementing analytical database for generating complex reports and preparing data for SSAS data initialization;

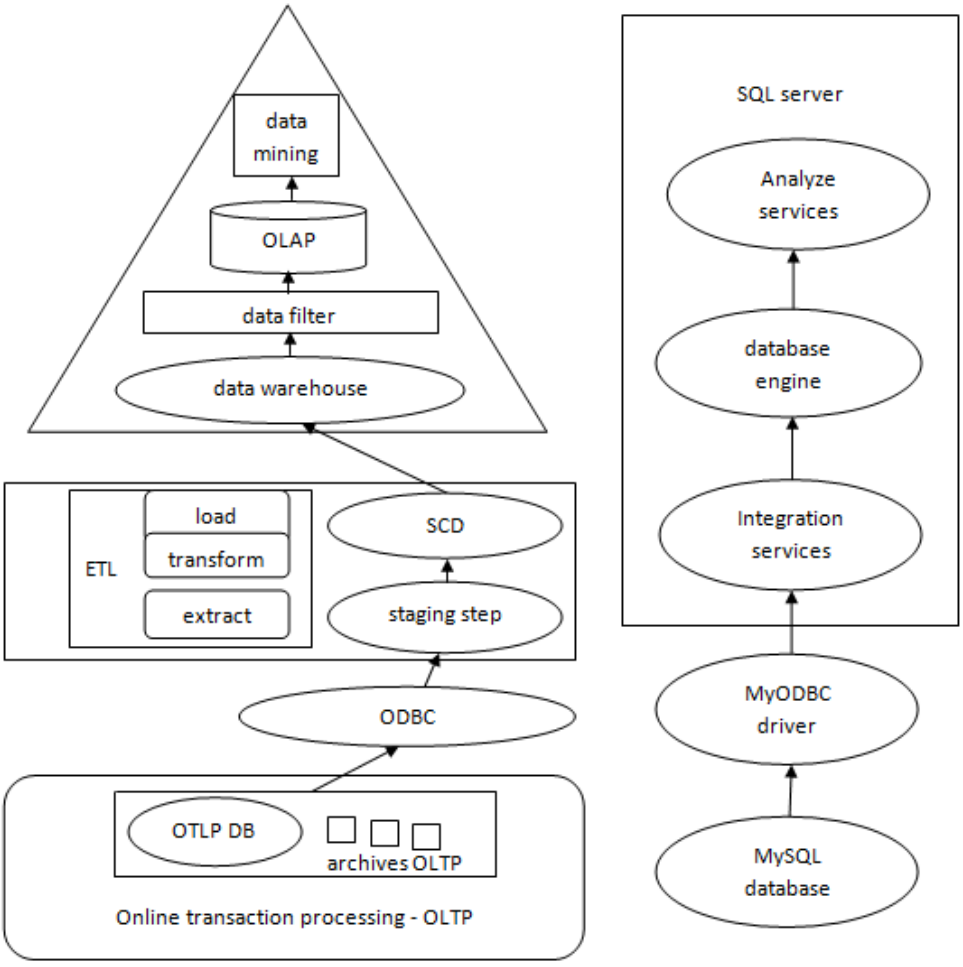


**Fig. 3.** DSS framework tools

### 4.4 Application components

Any analysis model can be configured directly from structures interrogating in the integrating services projects. Obvious method combines both extract and transform steps in the same flow and load the resulted data in the final destination model. It can create some data redundancies and replicate some common functionalities. Top-down methodology involves breaking down of a system to get insight into its compositional sub-systems.

It assure that any additional functionalities can be added with less effort. Staging method avoids any possibility of replicating common functionalities by using some intermediary steps. figure 2 shows the difference between the staging and obvious method. The components of staging methods as it seen in the figure 2 can be summarized as follow:

- *Extraxt1 & load1*: includes copying data in the intermediary database without change its structures or values;
- *Extract2 & trasform2 & load2*: extract any new data instances in staging database to load it in the data warehouse. This stage can include data transformation to adapt the data to the requirements of data warehouse;
- *Extract3 & transform3 & load3*: Extracted data from data warehouse are filtered and transformed to meet the requirement of analysis system;

### 4.5 Data source integration

Backup data file represents a copy of database. It is generated in certain time and has the same data as operational database (ODB) at that period. This content also includes erroneous data that exists at ODB. These type of files can be used in case of inexistent a system for data archiving and enable recovering data in case of any system failure. OTLP(online transaction processing) or ODB is a database that is designed and normalized to avoid redundant data and facilitate the operations (insertion, deletion and updating) captured from transactional system, it is used to store data using the relational database technology to ensure the integrity. It dedicates to serve the transactional systems so it just include the current activity without historical data.


**Fig. 4.** Recovering historical data using backup data files

figure 4 shows components used for building data warehouse for our case study. OBD and three backup data file. It shows data content for every source and its period. The grey portion indicates that data is partial and not complete. Older backup file is used for recovering data of the grey zone. The example presented in the figure 4 shows that the entire historical data can't be recovered so it remains some periods where data is not complete. Some procedures should be taken for dealing with this situation. ETL is used to update the data warehouse by capturing changes data from transactional system and loading them in data warehouse. The process of elimination method is used in SSIS by the component slowly changing dimension (SCD) that is shown in figure 3. It is dealing with insert and update commands properly but it does not process the delete command so all the deleted data captured from backup data files are migrated to data warehouse. Most of deleted records represent historical data and should be kept in data warehouse. The deleted records can be classified as following:

- *Missed historical data*: are valid data that should be kept in the data warehouse to serve the analytical purposes. Missed historical data represent the instances of historical data that cannot be recovered or not completed;
- *Cancelled data*: present removed data from the transactional system that should be identified and isolated. These records were removed from the transactional system because of

correcting users mistakes, duplicate insert, eliminate transaction such as an cancelled selling transaction. We can mention to these records as cancelled data;

### 4.6 Data quality problem

Last step of designing data warehouse is to define the rules used for data validation. Many Erroneous data can be revealed at stage of data integration. Determining the reason of appearing erroneous data can ease the task of cleansing them. By analytical point of view, a classification of erroneous data is proposed as following:

- *Platform error*: any erroneous data generated by system without intervention of users can be considered as system error, the reason of existing such these erroneous data is due to platform reliability;
- *System error*: error generated by transactional system due to problems of modelling or programming the system;
- *Users mistakes*: mistakes that are generated and not affect the transactional application. such these mistakes should be identified cause they can affected the analytical database. An example of these errors is reversing the dimensions values of length and width. This reversing is not affect the total quantity of materials but it can affect any clustering or classification method;
- *Cancelled data*: mistakes records that registered in system and removed later, the gap time between inserted and deleted operation enable to some of these records to be captured and inserted in data warehouse at the integration data step;

Some validation data rules have been defined to capture and isolate these data. The validation rules are classifies in two categories, unconditional and conditional validation rule. Unconditional validation rule: where there are two related table, validation of the first table is dependent on the second table. Unconditional validation means that no need of pre validation for the second table. We mentioned the second table as validation table. In case conditional validation rule, the data of validation entity should be verified and approved before using it in validation another entity. This type of rule requires some pre-validation. Different rules require different pre-validation steps. Thus, a flow that shows dependencies between rules is important to indicate the order in which these rules should be executed.

## 5 Conclusions

Business Intelligence is concept that include a set of techniques and methods that aim to configure high level tool that served the analytical purposes in order to support the decision maker. Tools of business intelligence become more variant and in hand for every possible user. So it is time for medium companies to get the benefit from such solution. Medium companies have a large data amount that need to be analyzed but it cannot invest a lot. At this paper, we try to build a decision support system tailored for timber export company. Prototyping methodology is used as it suitable for innovative projects where no previous examples of this type exist. There are many varies architectures for development DSS. Our proposed system is divided into small and isolated tasks or components. This enable every task to be smaller and specific but the system is more flexible for maintenance and evolve. SQL server services (SAS) is one of largest companies that also offers BI solutions addressed to medium companies. SAS has a bunch of predefined functions and layouts that decrease programming effort, so implementing the prototyping methodology becomes easier. Building data warehouse requires integration and validating data from different sources. This task includes several steps or processes such as extact, transform, filtering, cleaning and load. Data cleansing is one of most difficult tasks.

Different approaches and techniques are proposed for validation and cleansing data from single or multiple sources.
In this paper, It was solved the problem of integration from single source but by using backup data files and not archiving system. The accuracy of integrated data can not be guareented 100% especially when the legacy data is not complete. The purpose of data warehouse is to serve applications dedicates to solve analytical problem. Thus, it is important to adopt solution that balanced between the quality and cost. Automat solution is prefered in data cleansing than manual working. In our case stude, we apply some tehniques for filtering and cleaning the historical data for backup data files problem.

**Acknowledgment**
Part of this work is done under the auspices of the doctoral studies within the Doctoral School of Economic Informatics, Bucharest University of Economic Studies.

**References**

[1] A. Sohollo, "Using Business Intelligence in IT Governance Decision Making," in *Governance and Sustainability in IS*, M. Nüttgens, A. Gadatsch, K. Kautz, I. Schirmer and N. Blinn, Eds. Germany: Springer, IFIP AICT 366, IFIP International Federation for Information Processing 2011, Part 1, pp. 3–15.

[2] J.A. O'Brien and G.M. Marakas, *Management Information Systems*, 10th ed. New York, U.S.A.: The McGraw-Hill Companies, 2011.

[3] R.L. Sallam, J. Richardson, J. Hagerty and B. Hostmann. (January 2011). *Magic Quadrant for Business Intelligence Platforms*. Available: http://www.board.com/download/press/EN/Gartner_BI_MagicQuadrant_2011.pdf

[4] Gartner. (September 2012). *Business Intelligence, Mobile and Cloud Top the Technology Priority List for CIOs in Asia: Gartner Executive Programs Survey*, Available: http://www.gartner.com/it/page.jsp?id=2159315.

[5] Gartner. (April 2012a). *Gartner Says Worldwide Business Intelligence, Analytics and Performance Management Software Market Surpassed the $12 Billion Mark in 2011*, Available: http://www.gartner.com/it/page.jsp?id=1971516.

[6] H.J. Watson, C. Fuller and T. Ariyachandra, "Data warehouse governance: best practices at Blue Cross and Blue Shield of North Carolina," *Decision Support Systems*, Vol. 38(3), pp. 435-450, 2004.

[7] T. Ariyachandra, H. J. Watson, "Key organizational factors in data warehouse architecture selection", *Decision Support Systems 49*, 2010, 200–212.

[8] T. R. Sahama, P. R. Croll, "A Data Warehouse Architecture for Clinical Data Warehousing", in Roddick, J. F. and Warren, J. R., Eds. *Proceedings Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007) CRPIT*, 68, pages pp. 227-232, Ballarat, Victoria.

[9] B.A. Devlin, P.T. Murphy, "An architecture for a business and information system," *IBM Systems Journal 27 (1)* (1988) 60 – 80.

[10] W.H. Inmon., "DW 2.0 Architecture for the Next Generation of Data Warehousing", *DM Review*, Apr 2006, Vol. 16 Issue 4, p.8-25.

[11] W.H. Inmon, "Building the Data Warehouse", *Third Edition, York: John Wiley & Sons*, 2002.

[12] H.-G. Hwang, C.-Y. Ku, D. Yen and C.-C. Cheng, "Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in Taiwan," *Decision Support Systems 37.1* (2004): 1-21.

[13] Nilakanta, Sree, K. Scheibe, and A. Rai. "Dimensional issues in agricultural data warehouse designs,"

*Computers and electronics in agriculture 60.2* (2008): 263-278.
[14] W.H. Inmon, "Building the Data Warehouse," *Wiley*, New York, 1996.
[15] R. Kimball, "The Data Warehouse Toolkit", *Wiley*, New York, 1996.
[16] C. BOJA, A. POCOVNICU and L. BĂTĂGAN, *Distributed Parallel Architecture for "Big Data",* Informatica Economică, Vol. 16, 2012.
[17] Q. Chen, U. Dayal and M. Hsu, *A Distributed OLAP Infrastructure for E-Commerce,* HP Laboratories, Palo Alto, California, USA.
[18] R. Davenport, *ETL vs ELT,* Insource IT Consultancy, Insource Data Academy, June 2008.
[19] E. Rahm and H. H. Do, *Data Cleaning: Problems and Current Approaches,* IEEE Data Engineering Bulletin, Vol. 23, 2000.
[20] V. Raman and J. M. Hellerstein, *Potter's Wheel: An Interactive Data Cleaning System,* 2001.
[21] H. H. Do and E. Rahm, *On Metadata Interoperability in Data Warehouse,* Techn. Report, Dept. of Computer Science, http://dc-pubs.dbs.uni-leipzig.de/files/Rahm2000DataCleaningProblemsand.pdf [May 3, 2016].
[22] S. F. Liu and Y. Lin, 2010, *Grey Information. Theory and practical Applications,* Springer-Verlag, London, 2010.
[23] L. Cabibbo and R. Torlone, *Querying multidimensional databases*, Proceedings of the 6th DBLP Workshop, pages 253–269, 1997.
[24] F. Dehne, T. Eavis, S. Hambrusch and A. Rau-Chaplin, *Parallelizing the datacube*, International Conference on Database Theory, 2001.
[25] C. Diaconu, C. Freedman, E. Ismert, P.-A. Larson, P. Mittal and R. Stonecipher, N. Verma, and M. Zwilling. Hekaton, *SQL server's memory-optimized OLTP engine*, Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD '13, pages 1243–1254, New York, NY, USA, ACM, 2013.
[26] M. L. DESPA, *Comparative study on software development methodologies*, Database Systems Journal vol. V, no. 3/2014.
[27] J. E. Cooling, T. S. Hughes, "The emergence of rapid prototyping as a realtime software development tool", *Proceedings of the Second International Conference on Software Engineering for Real Time Systems*, 18-20 Sep. 1989, Cirencester, UK, Publisher: IET, 1989, pg. 60-64.

**Yasser AL-HADAD** has graduated the Faculty of Information Management at the Romanian American University in 2008. He received a master's degree in Economic Informatics from Romanian American University in 2010. Since then he is a PhD candidate, studying to obtain his PhD in the field of economic informatics.

**Răzvan Daniel ZOTA** has graduated the Faculty of Mathematics – Computer Science Section at the University of Bucharest in 1992. He has also a Bachelor degree in Economics, a postgraduate degree in Management from SNSPA Bucharest, Romania. In 2000 he has received the PhD title from the Academy of Economic Studies in the field of Cybernetics and Economic Informatics. From 2010 he is supervising PhD thesis in the field of Economic Informatics.