

WWW.JAQM.RO

**JOURNAL
OF
APPLIED
QUANTITATIVE
METHODS**

Knowledge Dynamics

**Vol. 5
No. 1
Spring
2010**

ISSN 1842-4562

JAQM Editorial Board

Editors

Ion Ivan, University of Economics, Romania

Claudiu Herteliu, University of Economics, Romania

Gheorghe Nosca, Association for Development through Science and Education, Romania

Editorial Team

Cristian Amancei, University of Economics, Romania

Catalin Boja, University of Economics, Romania

Radu Chirvasuta, University of Medicine Carol Davila, Romania

Irina Maria Dragan, University of Economics, Romania

Eugen Dumitrascu, Craiova University, Romania

Matthew Elbeck, Troy University, Dothan, USA

Nicu Enescu, Craiova University, Romania

Bogdan Vasile Ileanu, University of Economics, Romania

Miruna Mazurencu Marinescu, University of Economics, Romania

Daniel Traian Pele, University of Economics, Romania

Ciprian Costin Popescu, University of Economics, Romania

Aura Popa, University of Economics, Romania

Marius Popa, University of Economics, Romania

Mihai Sacala, University of Economics, Romania

Cristian Toma, University of Economics, Romania

Erika Tusa, University of Economics, Romania

Adrian Visoiu, University of Economics, Romania

Manuscript Editor

Lucian Naie, SDL Tridion

JAQM Advisory Board

- Luigi D'Ambra**, University "Federico II" of Naples, Italy
Ioan Andone, Al. Ioan Cuza University, Romania
Kim Viborg Andersen, Copenhagen Business School, Denmark
Tudorel Andrei, University of Economics, Romania
Gabriel Badescu, Babes-Bolyai University, Romania
Catalin Balescu, National University of Arts, Romania
Avner Ben-Yair, Sami Shamoon Academic College of Engineering, Israel
Constanta Bodea, University of Economics, Romania
Ion Bolun, Academy of Economic Studies of Moldova
Recep Boztemur, Middle East Technical University Ankara, Turkey
Constantin Bratianu, University of Economics, Romania
Irinel Burloiu, Intel Romania
Ilie Costas, Academy of Economic Studies of Moldova
Valentin Cristea, University Politehnica of Bucharest, Romania
Marian-Pompiliu Cristescu, Lucian Blaga University, Romania
Victor Croitoru, University Politehnica of Bucharest, Romania
Cristian Pop Eleches, Columbia University, USA
Michele Gallo, University of Naples L'Orientale, Italy
Angel Garrido, National University of Distance Learning (UNED), Spain
Bogdan Ghilic Micu, University of Economics, Romania
Anatol Godonoaga, Academy of Economic Studies of Moldova
Alexandru Isaic-Maniu, University of Economics, Romania
Ion Ivan, University of Economics, Romania
Radu Macovei, University of Medicine Carol Davila, Romania
Dumitru Marin, University of Economics, Romania
Dumitru Matis, Babes-Bolyai University, Romania
Adrian Mihalache, University Politehnica of Bucharest, Romania
Constantin Mitrut, University of Economics, Romania
Mihaela Muntean, Western University Timisoara, Romania
Ioan Neacsu, University of Bucharest, Romania
Peter Nijkamp, Free University De Boelelaan, The Netherlands
Stefan Nitchi, Babes-Bolyai University, Romania
Gheorghe Nosca, Association for Development through Science and Education, Romania
Dumitru Oprea, Al. Ioan Cuza University, Romania
Adriean Parlog, National Defense University, Bucharest, Romania
Victor Valeriu Patriciu, Military Technical Academy, Romania
Perran Penrose, Independent, Connected with Harvard University, USA and London University, UK
Dan Petrovici, Kent University, UK
Victor Ploae, Ovidius University, Romania
Gabriel Popescu, University of Economics, Romania
Mihai Roman, University of Economics, Romania
Ion Gh. Rosca, University of Economics, Romania
Gheorghe Sabau, University of Economics, Romania
Radu Serban, University of Economics, Romania
Satish Chand Sharma, Janta Vedic College, Baraut, India
Ion Smeureanu, University of Economics, Romania
Ilie Tamas, University of Economics, Romania
Nicolae Tapus, University Politehnica of Bucharest, Romania
Timothy Kheng Guan Teo, National Institute of Education, Singapore
Daniel Teodorescu, Emory University, USA
Dumitru Todoroi, Academy of Economic Studies of Moldova
Nicolae Tomai, Babes-Bolyai University, Romania
Victor Voicu, University of Medicine Carol Davila, Romania
Vergil Voineagu, University of Economics, Romania



	Page
Knowledge Dynamics	
Ion SMEUREANU, Andreea DIOSTEANU, Liviu Adrian COTFAS Knowlegde Dynamics in Semantic Web Service Composition for Supply Chain Management Applications	1
Ion IVAN, Cristian CIUREA, Sorin PAVEL Very Large Data Volumes Analysis of Collaborative Systems with Finite Number of States	14
Mladen CUDANOV, Ondrej JASKO, Gheorghe SAVOIU Interrelationships of Organization Size, Information and Communication Technology Adoption	29
Cezar VASILESCU Modeling the Reliability of Information Management Systems based on Mission Specific Tools Set Software	41
Hemanta SAIKIA, Dibyojyoti BHATTACHARJEE Data Mining into the Websites of Management Institutes using Binary Representation	53
Quantitative Methods Inquires	
Ida CAMMINATIELLO, Luigi D'AMBRA Visualization of the Significant Explicative Categories using Catanova Method and Non-Symmetrical Correspondence Analysis for Evaluation of Passenger Satisfaction	64
Alexandru SMEUREANU, Stefan Daniel DUMITRESCU Implementing a GIS application for Network Management	73
Vandna JOWAHEER, Varunah LALBAHADOOR, Roshan RAMESSUR, Lutchmee DOSORUTH A Multifactor Statistical Model for Analysing the Physico-Chemical Variables in the Coastal Area at St-Louis and Tamarin, Mauritius	89
Ion DOBRE, Adriana AnaMaria ALEXANDRU, Octavia LEPAS The USA Shadow Economy and the Unemployment Rate: Granger Causality Results	98
Michael A. LEWIS The Physics of Inflation: Newton's Law of Cooling and the Consumer Price Index	105
Mihaela DRAGOTA, Natalia SUSANU Methods of Portfolio Management for Listed Shares. Some Features for the Romanian Private Pension Funds	113

	Page
Applied Quantitative Methods in Medicine	
Jeremy R. SCHWARTZ, Shlomo MARK, I. YAAR, S. MORDECHAI SPECTRALYZER: A Comprehensive Program to Classify FTIR Microscopic Data Applied for Early Detection of Critical Ailments	133
Himanshu PANDEY, Kamlesh Kumar SHUKLA The Probability Model for Risk of Vulnerability to STDs/or HIV Infection among Pre-Marital Female Migrants in Urban India	145
Alexandru-Ionut PETRISOR, Liviu DRAGOMIRESCU, Cristian PANAITIE Alexandru SCAFA-UDRISTE, Anamaria BURG Subject-Level Trend Analysis in Clinical Trials	152
Salah UDDIN, Arif ULLAH, NAJMA, Muhammad IQBAL Statistical Modeling of the Incidence of Breast Cancer in NWFP, Pakistan	159
C. PONNURAJA An Empirical Investigations of Meta-Analysis using Randomized Controlled Clinical Trials in a Particular Centre	166
Book Reviews	
Aura POPA Book Review on REGRESSION MODELING: METHODS, THEORY, AND COMPUTATION WITH SAS, by Michael J. PANIK, Chapman&Hall/CRC, Taylor&Francis Group, Boca Raton, FL, USA, 2009	176
Cristian CIUREA Book Review on DEVELOPING RIA APPLICATIONS ("DEZVOLTAREA APLICATIILOR RIA"), by Andrei TOMA, Radu CONSTANTINESCU, Mihai PRICOPE, Floarea NASTASE, ASE Publishing House, Bucharest, 2010	180

KNOWLEDGE DYNAMICS IN SEMANTIC WEB SERVICE COMPOSITION FOR SUPPLY CHAIN MANAGEMENT APPLICATIONS¹

Ion SMEUREANU²

PhD, University Professor, Department of Economic Informatics,
Dean of Faculty of Cybernetics, Statistics and Economic Informatics,
University of Economics, Bucharest, Romania

E-mail: smeurean@ase.ro



Andreea DIOSTEANU³

PhD Candidate, University Instructor, Department of Economic Informatics,
University of Economics, Bucharest, Romania

E-mail: andreea.dioşteanu@ie.ase.ro



Liviu Adrian COTFAS⁴

PhD Candidate, University Instructor, Department of Economic Informatics,
University of Economics, Bucharest, Romania

E-mail: liviu.cotfas@ase.ro



Abstract: *Semantic Web Service technology can play a vital role in today's changing economic conditions, as it allows business to quickly adapt to market changes. By combining individual services into more complex systems, web service composition facilitates knowledge dynamics and knowledge sharing between business partners. The proposed framework uses a semi-automatic approach as manual web service composition is both time-consuming and error-prone.*

Key words: *semantic web service composition; multi agent systems; fractal theory; location based application; supply chain management*

1. Introduction

Nowadays, the enterprise environment is heavily influenced by the evolution of the IT&C technologies. These changes mainly influence the collaborative economic environment the dynamics of knowledge management and increase the level of competition at a global scale. Therefore, adapting and efficiently taking advantage of these changes represents a real challenge for the top and medium management. The enterprise steering committee has

to permanently by updated with the latest discoveries so that to obtain maximum satisfaction.

The 21st century, knowledge management has changed a lot due to software and hardware progress. As a consequence, we can affirm that one of the main features of knowledge management is its dynamism. As part of the enterprise environment, knowledge represents added value and contributes as a key factor in assuring sustainable economic growth and increased profit levels. An intelligent management strategy has to be aware of all these changes and to take action so that to maximize the effective use of knowledge. In order for them to gain market shares and to be competitive it is very important to have efficient knowledge management strategies and to be able to take advantage of the new technologies that are continuously emerging. Furthermore, because companies' profitability is mainly influenced by the manner in which their products or services are perceived by customer we can state the fact that the growth and implicitly the investment strategies become product and customer oriented structures. In the context of dynamic business, maximizing and optimizing business performance is a critical requirement for profitability.

The modern economic behaviors are the most accurate and visible proof of the impact the knowledge dynamics has over the traditional types of needs and opportunities. Moreover, all these transformations that occurred are closely connected and have a great influence over the globalization trend the economy is heading towards. The global market can be characterized by increased levels of interoperability that reflect into the integration of multiple and different information systems which are able to share, manipulate and combine knowledge so that to facilitate the enterprise (or generally speaking organizations) collaboration process.

Having given all the above facts, we can conclude that the very frequently used term "integrated company" is not actual any more. Presently, it is substituted by collaborative information systems composed of business networks that are linked to independent partners that provide individual services and goods.

Knowledge management in such complex business and information structures can be defined as an approach, strategically targeted, so that to motivate the members of an organization to develop and use their cognitive capacities, sources of information, experience and abilities by subordinating their own objectives to the overall objectives. In the organizational environment, knowledge is derived from the information that is processed by those who have the capacity to effective action, by assimilation and mainstream understanding, followed by operationalizing the given contexts (Dragomirescu, 2001).

This paper presents a software agent based framework for modeling and enhancing the performance of knowledge management and increasing its level of dynamism by facilitating enterprise interoperability with the help of an automatic web service composition module.

The first section of the article consists of a short literature review so that to establish the place of our research in the current international research trends. And also draws a parallel among different automatic semantic web service composition solutions. In the future sections we enlarge upon the technologies used for implementing the solution, the architecture of the proposed framework. Furthermore, we will present a case study application that we developed to validate the agent based search module of our architecture.

2. Current SWC solutions Analysis

In order to implement interoperable and collaborative applications, the current research trend is heading towards determining the most suitable model for automatic or semi-automatic composition of web services. There are many proposed software solutions architectures for web service compositions, however up to now none of them has been fully implemented in the real business world due to various drawbacks:

- lack of service availability and accessibility;
- large data volumes caused by numerous service descriptions;
- increased search time in web service directory;

The syntactic description of web service is not sufficient for the automation of the web service composition and discovery process. A more detailed semantic description which can annotate the request on the basis of common semantics is needed.

Collaborative and knowledge based applications are to be composed of a set of semantically annotated web services and no longer be implemented separately. Therefore, the composition of semantic web services has enormous potential in improving the integration and collaboration in a wide variety of applications: business-to-business, location based services, supply chain, etc.

Semantic web service composition mainly consists of designing and implementing complex business logic workflows by organizing semantic annotated web services so that to obtain a single semantic complex web service. In order to enable composition, web service functionalities have to be described in detail by using either semantic or functional annotation. In this manner extra information about their main function or about the way the services behave is provided different from the classical syntactic web service annotation. For further information regarding the functionality of a service such as preconditions, conditions, effects, etc. the RDF description language is used. The RDF format uses semantic elements of ontologies that have to be previously established depending on the application type.

By composing services not only that certain components may be used, but processes can be integrated into applications by modeling the business flow. Automated web services composition techniques can be classified into two categories: static and dynamic. (Chakraborty and Joshi 2001).

The static web service composition technique refers to the fact that business flow designers or business analysts manually implement the composition by predefining business processes and by describing the interaction of their web services' components. Dynamic composition of services is not based on such predefined processes. It is rather based on existing web service retrieval and dynamic assembly in order to meet the initial demanding on the semantic content used in their annotation.

The traditional means for describing web services' functionality do not have enough semantic information to be used in the composition. The new trend in web services development can be characterized by the use of ontologies in order to achieve a complete description of them. This type of semantic annotated web services is called semantic web services (Mcilraith, Son and Zeng 2001). The semantic web services composition problem also focuses on the automatic and flexible discovery of services.

The recent research in semantic web services are largely focused on handling requests for task-oriented services and on obtaining information in distributed environments such as the internet information.

One of the first papers (Zhang, et al. 2008) which tackle with this subject performs a detailed analysis to determine the integrated applications' pattern for modeling business processes within enterprises and introduces a semi-dynamic module for automatic composition of such semantic web services. In order to achieve this objective, the authors use an integrated applications' ontology in the field of modeling business processes within enterprises to provide not only basic semantic concepts, but also concepts and terminologies that are used to describe services and to define abstract business processes.

Abstract business processes represent the static part of the semi-dynamic semantic web services composition and it is dynamically attested at runtime through the automatic discovery process. Such a method combines the advantages of static and dynamic composition of services and carries out the semantic based integration for business processes modeling applications. One of the major problems the automatic service composition faces is related to assembling individual services based on their functional specifications to create added value and satisfy a particular service request.

Most approaches to automatic composition of services are based on artificial intelligence planning techniques (Thakkar, et al. 2002), (Mcilraith and Son, 2002), (Wu, et al. 2003). All these techniques require that all relevant service descriptions are loaded into a reasoning engine. Due to the very large number of service descriptions and the weak link between suppliers and consumers, services are indexed in specific directory. The problem occurs when loading a directory that contains such a large volume of data. To solve this problem it is necessary that the planning algorithms should be changed so that only the relevant descriptions to be dynamically extracted from the directory during the composition process (Constantinescu 2004), (Constantinescu, 2005).

In such an approach, a single service composition involves many complex queries of the associated indexed directory. For example, if composition algorithms such as "forward chaining" are being used, each query processing up to 20% of data stored in the directory even if it has an optimized index structure (Constantinescu, 2005). Taking into account that such a directory is a public resource where bottlenecks may occur the problem becomes more complex and complicated to handle.

In addition to the above presented solutions there are some abstract ones based on Colored –Petri Nets (Qian, Lu and Xie 2007), Petri Nets (Hamadi and Benatallah 2003), mathematical models, etc.

In (Qian, Lu and Xie 2007), the authors focus on automatically synthesizing desired composite service through available services. The first contribution of this paper is a general description model of services. This paper also proposes a Colored Petri Net (CPN) based service model MOAP which indicates the relationship between messages (data) and service behaviors. The second contribution of the paper is an effective technique for automatic service composition.

In (Hamadi and Benatallah 2003) the authors propose a Petri net-based algebra for composing Web services. The operators that are used for composing web services are directly link to Petri nets by expressing their semantics in terms of this particular type of nets. The first result of such link is the fact that every service can be expressed as a Petri Net. The advantage brought up by the use of Petri Nets is related to their ability to simulate workflow patterns, operations and properties of web service composition.

The non-abstract web service composition techniques can be grouped into several guidelines according to the research trends that they follow.

Table 1. Semantic web service composition techniques and their objectives

Web service Composition technique	Objective
<i>AI-planning</i>	Finding a path (action plan, sequence of actions) from the initial state to a preset state (target state).
<i>Chaining techniques</i>	Finding dependencies between services in so that to synthesize a composition plan that matches the request

In (Talantikite, Aissani and Boudjlida 2008) we can see a complex approach to web service composition that combines graph based abstract models, chaining algorithm of expert system and semantic annotations. The authors use an inter-connected network of semantic web services that are semantically annotated with OWL-S. The concepts of the ontology enable them to measure the similarity between the outputs and inputs. The advantage brought by this approach is that the composition algorithm can find several composition plans in only one exploration phase. Afterwards the composition plans are being filtered according to precise quality criteria composed of: similarity, time and memory space quality metrics.

3. Technologies Used for Developing the Business Application Framework

3.1. Semantic Web Services

Web Services are internet based technologies that enable the process of making connections. Services represent the elements that are being connected with the help of web services. A service is the endpoint of a connection. Moreover, a service has some type of underlying computer system that supports the connection offered. The combination of services – both internal and external to an organization determines a service-oriented architecture.

Services are platform-independent software entities that can be described, published, discovered, and loosely coupled in different ways. They can perform different and complex functionalities from answering simple requests to executing sophisticated business processes requiring peer-to-peer relationships among multiple layers of service consumers and providers. Furthermore, they have features that permit software reengineering and also software reusability and transformation into network-available service.

Presently, the most highly used web services technologies only provide syntactic descriptions, making it difficult for requesters and providers to interpret or represent statements such as the meaning of inputs and outputs or applicable constraints.

As it was stated in the first section, enterprise interoperability and knowledge dynamics represent some of the most important guidelines in developing efficient business flow modeling application. Due to this fact, a lot of research was carried on in this domain. The current trend is to determine the most suitable applications' architectures that fulfill the previously mentioned requirements.

Such applications should be characterized by:

- an increased level of platform-independence for each component;
- interoperability;
- scalability ;

- flexibility;

Web services are software applications that are characterized by these features and are available in the distributed environment of the Internet and are based on XML (eXtensible Markup Language). The functionality of web services is usually presented in a syntactic manner by using standards like: UDDI (Universal Description, Discovery and Integration), WSDL (Web Service Description Language), SOAP (Simple Object Access Protocol).

UDDI is a virtual registry that exposes information about Web services.

WSDL provides a XML based model format for describing web services functionality. WSDL represents a syntactic interface for web services. The description provided by WSDL is separated into two parts: an abstract description of functionality and a particular one that illustrates the details of the web service ("how" certain behaviors are being implemented and "where). WSDL describes only the syntactic interface of Web services.

SOAP is an XML based protocol that is used to exchange structured information in a decentralized and distributed environment. The role of the XML is to define an extensible framework of messages. With the help of such a framework messages, that are to be exchanged through other related protocols, are being constructed. SOAP is not related to any particular programming model and is independent from any specific semantics (Gudgin, et al. 2003).

However, the syntactic description does not provide all the necessary information about a service especially for the web service composition process. For this reason, the semantic annotated web services are used. In order to achieve this objective, they provide functionalities for creating, managing and transforming semantic mark up. The semantic annotations are considered to be conditions and effects of web services and are explicitly declared in the Resource Description Format (RDF) using terms from pre-agreed ontologies. Consequently, it enhances the ability of smart agents to understand, transform and deliver messages over the web.

3.2. Semantic Web Service Composition

The main goal for using semantic web services is that to increase the possibility of automated service discovery, composition invocation and monitoring over the web and in this manner to assure interoperability and collaboration between different business flow modeling applications.

The semantic web community is conducting complex studies in order to determine the most efficient methods for synthesizing web services' complex behaviors and determining an universal semantic representation for the transformation that occur. In this way, web services will be easier to discover on the internet by specific subscribers and also it will be easier to assure web service composition and interaction in Service Oriented Architecture based solutions.

The applications that implement web service composition have a SOC (Service Oriented Computing) architecture that is based on SOA. This type of architecture uses services to support the development of rapid, low-cost, interoperable, evolvable and distributed applications.

As it is stated in (Papazoglou, et al. 2007), according to SOC research road map, SOC provides a logical separation of functionality into three planes:

- service foundations at the bottom

- service composition
- service management and monitoring
 - This logical stratification is based on the need to separate:
- basic service capabilities provided by a middleware infrastructure and conventional SOA from more advanced service functionality needed for dynamically composing services,
- business services from systems-centered services
- service composition from service management.

3.3. Multi-agent Systems

Multi-agent systems are formed by a certain number of agents that interact with each other. The major advantage of such systems consists of the fact that simple individual behaviors combine into complex ones. Furthermore, another important feature of multi agent systems refers to their ability of decomposing complex problems into more easily manageable sub problems. According to (Bodea and Mogoș 2007), this idea can be applied in many research domains such as for decomposing complex based geospatial problems or supply chain management problems: negotiations, discovery, offer analysis etc. Moreover, they can be used to complex data mining in logistic applications.

In (Genong, et al. 2009) it is illustrated that agents can communicate and interact with each other through ontology language which stands for communication languages. Agents can be managed and discovered through a centralized directory, peer-to-peer discovery, or hybrid mechanism. Agent mobility provides a mechanism to extend stabilities and sustainability of semantic web services in a distributed environment.

One of the major uses of multi-agent systems is related to manufacturing companies. In this case, agents are used to gather information over the web and to transform it into knowledge by adding value. However, if agents are used without combining with semantic web services technology, they fail to respond to the continuously changing business environment in the nowadays knowledge driven society. According to (Weiming, et al. 2007), the failure is associated to the fact that they function on a predefined agreement without being flexible. On the other hand, pure web-based technologies, including web services, cannot fulfill the needs of virtual enterprises applications, because: they do not offer the possibility to automatically discover corresponding services at run time. Also, web service description offers only a technical presentation of the features offered and not a semantic one. Last but not least the description of business processes and security features implemented by using web services are not very reliable because this domain is still in an incipient phase.

Intelligent software agents have been used in enterprise independent software systems integration process, not only to assure an approach for functional integration, but also to facilitate the use of business intelligence and collaboration among enterprises for their communication, interaction, cooperation, pro-activeness, and autonomous intelligent decision making.

In order to achieve the objectives of the current enterprise interoperability trend, we propose a framework that combines web services and software agents so that to provide an efficient service discovery, retrieval, composition and integration. This paper proposes an agent-based service-oriented integration architecture, where enterprise Web services are

dynamically discovered using agent behaviors and specific ontology for communication and retrieval.

Multi-agent systems are closely connected to web service technology because they represent interoperable, portable and distributed solutions. Agents and web services may be related in different ways: agents use web services, web services are in fact agents or agents are composed of, deployed as, and dynamically extended by web services (Martin, et al. 2005).

4. Web Service Composition Framework

The proposed framework combines the template-based and AI Planning web service chaining approaches. Based on the composition request and user experience, both automatic and semi-automatic web service composition can be used. In order to facilitate the composition process, web services are managed by intelligent agents. This transforms the RESTful web services into Stateful entities (OASIS 2005) that can maintain their state between calls. Using many intelligent agents, each following its own goal during the composition process helps distribute the composition effort and also avoids performance bottle-necks. All web-services must be previously annotated using OWL-S in order to identify their purpose. While WSDL (Booth and Canyang 2007) provides a syntactic description containing information regarding the structure of the input and output parameters, semantics provide a description of what the web service actually does.

4.1. Framework Components

Warping one or more web services as an intelligent agent is also known as agentification (Yang, et al. 2004). Not only web services, but also legacy applications can be “agentified”, thus reducing the costs of implementing the proposed framework by using existing software.

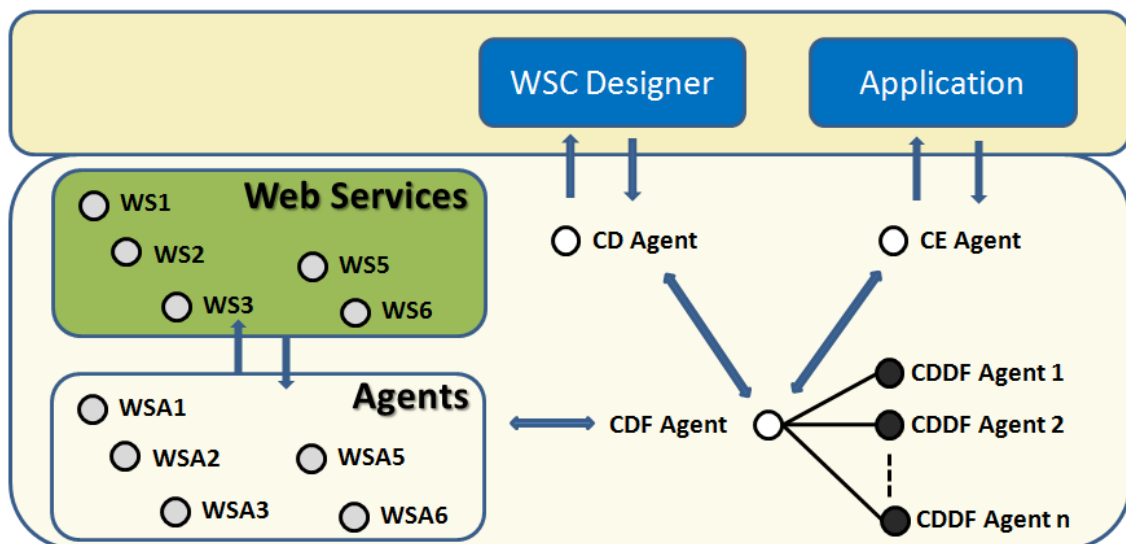


Figure 1. Web Service Composition Framework

The main components and relations between them can be seen in Figure 1.

Agents that can execute a specific task can be found using the Composition Directory Facilitator Agent (CDF) and the several domain specific CDDF. All agents in our

platform are registered with at least one CDDF agent. In order to find a web service capable of completing a specific task is need, the CDF agent will be first queried to identify the corresponding CDDFs, which in turn allow searching for agents. For a distributed implementation, multiple CDF agents can be run on different machines.

The Composition Design (CDA) and Composition Execution (CEA) Agents support the design and execution of web service chains. By assigning individual agents to each request, we both improve scalability and avoid creating a performance bottleneck.

The Web Service Composition (WSC) Designer allows creating and editing web service chains in an interactive manner. It also offers the possibility to test and validate the created web service chains as well as displaying several statistics regarding estimated duration and availability.

The communication between agents is implemented using FIPA compliant ACL messages. The proposed framework was implemented using JADE (Telecom Italia n.d.) (Java Agent Development Framework) and JENA (Hewlett-Packard n.d.) frameworks.

4.2. Fractal Web Service Composition

Web service composition implies finding a sequence of services that when called helps achieve a certain goal. In our approach, existing web service chains can be combined in a fractal like manner to easily create new and more complex web service chains. Similar to using components in classic programming, this approach has several advantages like the reduction of the development time and a higher QOS (Quality of Service) as a result of using already tested web services. Also, the solution becomes more adaptable, as, if needed, web services can be replaced with smaller effort.

Implementing a framework that eliminates the need to manually bind available services is particularly important in today's changing economic conditions, as it allows business using the proposed framework to quickly adapt to market changes without waiting for internal IT development projects or for long software vendor release cycles (Bussler 2007).

All web service chains in our framework are stored using WS-BPEL (Web Service Business Process Execution Language) (OASIS 2007) and can effectively be used as building blocks to create new service chains.

The design of the newly created chain starts in the WSC (Web Service Composition) Designer Module of our framework. The user has the possibility to either create a completely new chain or to modify an existing one. From the user's perspective the web service chain is composed from a succession of actions or sub-goals that must be performed in order to achieve a certain goal (Qian, Lu and Xie 2007). The user must specify the available input data and the requested output information.

For each action defined in the interface the following steps will be performed:

Step 1 – Action Matching: The CDA agent semantically queries the CDF and CDDF agents on all machines searching for Web Service Agents (WSA) that can perform the requested action and begins a parallel negotiation with all found agents. If no matching WSA can be found, the CDA agent will request the user to decompose the action into more elementary actions and will repeat the first step.

Step 2 – Input parameters: Every candidate WSA agent compares it's input data with the internally mapped web service or web service chain input parameters. In case any input parameters don't match, the WSA agent will itself query the CDF and CDDF agents in

order to find an agent or a chain of agents capable of performing the required transformation. If still needed the found agents can themselves repeat the procedure. Thus, the web service composition is performed in a manner similar to fractal theory, as each WSA agent can recursively call other agents. In order to limit the search space a maximum number of agents used to model an action can be specified.

Step 3 – Output parameters: Each caller agent, including the CDA verifies the correspondence between the called WSA web service output parameters and the requested output parameters. In case any output parameters don't match, an agent capable of performing the required transformation will be searched using the CDF and CDDF agents similar to Step 2.

Step 4: The best WSA agent or WSA agent chain is selected based on estimated execution time and availability statistics.

The final chain is stored so it can either be called directly or incorporated in new chains. A corresponding WSA is created and registered with the CDF and CDDF agents.

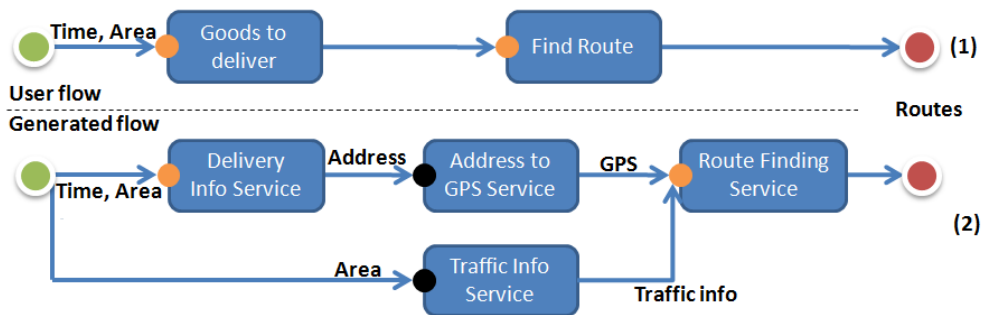


Figure 2. Web service composition using individual services

Figure 2 presents a composition scenario for location based services and supply chain systems in which the user's goal is to create a web service chain that finds the best route to deliver goods to customers. Based on his experience the user adds to actions: identifying the goods to deliver in a specific area at a specific time and finding a route (1). Based on this information, in Step 2, the CDA agent discovers the matching services "Delivery Info Service" and "Route Finding Service". "Address to GPS Service" is added to convert between "Delivery Info Service" output and "Route Finding Service" input type. "Traffic Info Service" supplies additional needed input data for the route finding service (2).

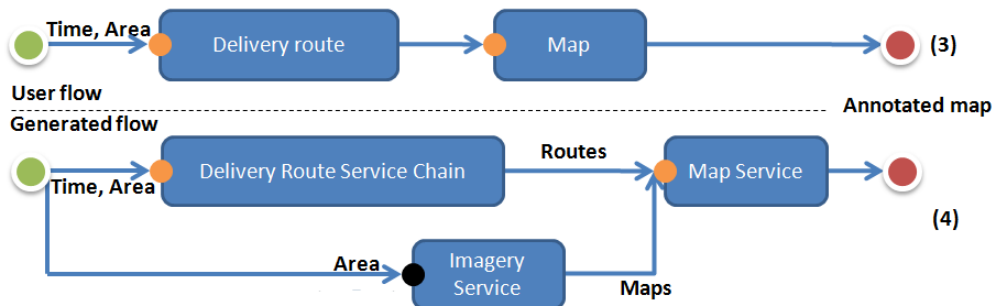


Figure 3. Web service composition using previously defined web service chains

In case the route should also be displayed on a map, a new web service chain can be created as shown in Figure 3. Based on users added actions (3), the framework selects

the previously defined chain and the "Map Service" capable of creating annotated maps based on information and map images. "Imagery Service" is automatically added to supply the parameters required by the Map Service.

Conclusions

Manual web service composition is both time-consuming and error-prone. The proposed framework allows the implementation of distributed semantic web service chains by using a semi-automatic approach. Organizations implementing such a solution will be able to better cooperate and share their knowledge. Moreover, such systems can easily be extended with new web service as they become available.

References

1. Bodea, C. and Mogos, R. **An Electronic Market Space Architecture Based On Intelligent Agents And Data Mining Technologies**, *Informatica Economica Journal* XI, no. 4, 2007, pp. 115-118
2. Booth, D. and Canyang, L. **Web Services Description Language (WSDL) Version 2.0 Part 0: Primer**, June 26, 2007, <http://www.w3.org/TR/2007/REC-wsdl20-primer-20070626> (accessed January 12, 2010)
3. Bussler, C. **The Fractal Nature of Web Services**, *Computer*, 40, 2007, pp. 93-95
4. Chakraborty, D. and Anupam, J. **Dynamic Service Composition: State-of-the-Art and Research**, Department of Computer Science and Electrical, Maryland University, Baltimore, 2001
5. Constantinescu, I. **Flexible and efficient matchmaking and ranking in service directories**, IEEE International Conference on Web Services (ICWS-2005), Florida: IEEE, 2005, pp. 5-12
6. Constantinescu, I. **Large scale, typecompatible service composition**, IEEE International Conference on Web Services (ICWS-2004), San Diego: IEEE, 2004, pp. 506-513
7. Dragomirescu, H. **Organizatii bazate pe cunoastere**, Research Institute for Artificial Intelligence, 2001, http://www.racai.ro/INFOSOCProject/Dragomirescu_st_g06_new.pdf (accessed March 2009)
8. Genong, Y., Liping, D., Wenli, Y., Peisheng, Z. and Peng, Y. **Multi-Agent Systems for Distributed Geospatial Modeling, Simulation and Computing**, in "Handbook of Research on Geoinformatics", Pennsylvania: Information Science Reference, 2009, pp. 196-205
9. Gudgin, M., Hadley, M., Mendelsohn, N. and Moreau, J.-J. **W3C Recommendation**, W3C Recommendation, June 2003, <http://www.w3.org/TR/soap12-part1/> (accessed January 2010)
10. Hamadi, R. and Benatallah, B. **A Petri net-based model for web service composition**, in "Proceedings of the fourteenth Australasian database conference", Adelaide, 2003, pp. 191-200
11. Martin, D., Burstein, M., McIlraith, S., Paolucci, M. and Sycara, K. **OWL-S and Agent-Based Systems**, In *Extending Web Services Technologies: The Use of Multi-Agent Approaches*, New York: Springer, 2005, pp. 53-77
12. McIlraith, S. and Son, C. T. **Adapting Golog for composition of semantic web services**, 8th International Conference on Principles and Knowledge Representation and Reasoning (KR-02), San Francisco: Morgan Kaufmann, 2002, pp. 482-496
13. McIlraith, S., Son, T. C. and Zen, H. **Semantic Web services**, *Intelligent Systems*, March 2001, pp. 46-53

14. Papazoglou, M. P. , Traverso, P., Dustdar, S. and Leymann, F. **Service-Oriented Computing: State of the Art and Research Challenges**, Computer 40, no. 11, November 2007, pp. 38-45
15. Qian, Z., Lu, S. L. and Xie, L. **Colored Petri Net Based Automatic Service Composition**, Asia-Pacific Service Computing Conference, The 2nd IEEE. Tsukuba Science City: IEEE, 2007, pp. 431-438
16. Talantikite, N. H., Aissani, D. and Boudjlida, N. **Semantic annotations for web services discovery and composition**, Computer Standards & Interfaces, 2008, pp. 1108-1117
17. Thakkar, S., Knoblock, C., Ambite, L. J. and Shahabi, C. **Dynamically composing web services from on-line sources**, AAI-2002 Workshop on Intelligent Service Integration, Edmonton, 2002, pp. 1-7
18. Weiming, S., Qi, H., Shuying, W., Yinsheng, L. and Hamada, G. **An agent-based service-oriented integration architecture for collaborative intelligent manufacturing**, Robotics and Computer-Integrated Manufacturing, no. 23 (2007), pp. 315-325
19. Wu, D., Parsia, B., Sirin, E. and Hen, J. **Automating DAML-S web services composition using SHOP2**, 2nd International Semantic Web Conference (ISWC-2003), Sanibel Island: Springer, 2003, pp. 195-210
20. Yang, H., Chen, F., Guo, H. and Xu, B. **Agentification for web services**, COMPSAC, New York, 2004, pp. 514-519
21. Zhang, K., Xu, R., Zhang, Y., Sai, Y. and Wang, X. **An Ontology Supported Semantic Web Service composition Method in Enterprise**, IEEE International Multi-symposiums on Computer and Computational, IEEE, 2008, pp. 222-227
22. * * * (W3C), World Wide Web Consortium, **OWL Web Ontology Language**, February 10, 2004, <http://www.w3.org/TR/owl-features> (accessed February 2010)
23. * * * Hewlett-Packard, **Jena Semantic Web Framework**, n.d. [http:// jena.sourceforge.net](http://jena.sourceforge.net) (accessed December 3, 2009)
24. * * * OASIS, **Web Services Business Process Execution Language Version 2.0**, April 2007, <http://docs.oasis-open.org/wsbpel/2.0/> (accessed January 8, 2010)
25. * * * Telecom Italia, **Java Agent Development Framework**, n.d. <http://jade.tilab.com/> (accessed October 19, 2009)
26. * * * OASIS, **Stateful Web Services**, March 7, 2005, <http://xml.coverpages.org/statefulWebServices.html> (accessed January 12, 2010)

¹ Acknowledgements

This article is a result of the project „Doctoral Program and PhD Students in the education research and innovation triangle“. This project is co funded by European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies.

² Ion SMEUREANU has graduated the Faculty of Planning and Economic Cybernetics in 1980, as promotion leader. He holds a PhD diploma in “Economic Cybernetics” from 1992 and has a remarkable didactic activity since 1984 when he joined the staff of Bucharest Academy of Economic Studies. Currently, he is a full Professor of Economic Informatics within the Department of Economic Informatics and the dean of the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies. He is the author of more than 16 books and an impressive number of articles. He was also project director or member in many important research projects. He was awarded the Nicolae Georgescu-Roegen diploma, the award for the entire research activity offered by the Romanian Statistics Society in 2007 and many others.

³ Andreea DIOSTEANU has graduated the Faculty of Economic Cybernetics, Statistics and Informatics in 2008 as promotion leader, with an average of 10. She is currently conducting research in Economic Informatics at Bucharest Academy of Economic Studies and she is also a pre-Assistant within the Department of Economic Informatics and .NET programmer at TotalSoft. During the bachelor years she participate in many student competitions both at national and international level obtaining a lot of first and second prizes. The most important competitions she was finalist in were Microsoft International Imagine Cup Competition, Software Design section (national finalist); Berkley University and IBM sponsored ICUBE competition where she qualified for the South Eastern Phase-Novatech. Furthermore, she also obtained the “N.N Constantinescu” excellence scholarship in 2007-2008 for the entire student research activity.



⁴ Liviu COTFAS is a Ph.D. student and a graduate of the Faculty of Cybernetics, Statistics and Economic Informatics. He is currently conducting research in Economic Informatics at Bucharest Academy of Economic Studies and he is also an Academic Preparator within the Department of Economic Informatics. Amongst his fields of interest are geographic information systems, genetic algorithms and web technologies.

VERY LARGE DATA VOLUMES ANALYSIS OF COLLABORATIVE SYSTEMS WITH FINITE NUMBER OF STATES¹

Ion IVAN²

PhD, University Professor, Department of Computer Science in Economics,
University of Economics, Bucharest, Romania

E-mail: ionivan@ase.ro ; **Web page:** <http://www.ionivan.ro>



Cristian CIUREA³

PhD Candidate, Department of Computer Science in Economics,
University of Economics, Bucharest, Romania

E-mail: cristian.ciurea@ie.ase.ro



Sorin PAVEL⁴

PhD Candidate, Department of Computer Science in Economics,
University of Economics, Bucharest, Romania

E-mail: pavelSORIN@gmail.com



Abstract: *The collaborative system with finite number of states is defined. A very large database is structured. Operations on large databases are identified. Repetitive procedures for collaborative systems operations are derived. The efficiency of such procedures is analyzed.*

Key words: *procedures, collaborative, state, database, operations, efficiency*

1. Collaborative systems with finite number of states

Within a collaborative system, a multitude of users or agents are involved in a distributed activity, most of the time being in several different places. In the large family of distributed applications, the collaborative system is identified by the common goal that the agent are working for and the great need of interaction that exists in the process of sharing and exchanging information and applications [1]⁵.

A software collaborative system is like a distribution company that seeks the increasing of sales. The difference between a collaborative system and a distributed one is given by the following attributes of the collaborative system:

- the system elements, both users and agents, interact with each other influencing the behavior of the system;
- the system components use shared resources in order to fulfill both their own objectives and their common goals;

- the system has permanent communication channels between users and agents;
- the agents' interests are not antagonistic (the agents have common and corporate interests).

The characteristics of collaborative systems are: complexity, reliability, portability, mentenability, structurability, stability, adaptability, operationability and integrability [3].

Types of collaborative systems counts:

- *collaborative systems in education*: active in the educational and research field and pursue increased performance and testing of the educational process;
- *collaborative systems of defense*: active in military field and are defined by strict rules of organizing and functioning;
- *collaborative systems in production*: pursuing increased production capabilities and product quality within distinct goods and services production units;
- *collaborative banking systems*: used by banks and financial units, these systems are analyzed along this paper in order to determine the parameters that influence the banking systems and all its components;
- *electronic business systems*: companies' departments are becoming more and more integrated, and clients are now users of e-business systems, thus replacing the traditional security mechanisms with authorization software – the modern security systems which mange and store users' data and correlate them with the access rules of the organization;
- *public administration systems*: used for managing tax collection, for integrated management of human resources and payroll, for querying city hall databases on citizen demand;
 - *media software development systems*: media applications development was indirectly caused by the increasing of common use electronic devices; these systems include commutations stations for wireless, terrestrial, satellite and cable infrastructure.

The agents of a collaborative system interact dynamically. Therefore, the system should be flexible in ability to execute transactions. Agents, servers, data warehouses and transactions are all elements which generally compound distributed systems, but the nature of transactions between agents and shared objectives of the agents are the main parts of a collaborative system.

Let S_1, S_2, \dots, S_n be the array of states that a collaborative system cover. Changing from S_i state to S_j state is done by a message, a command, a document d_{ij} . Table 1 contains the matrix of changing by documents from one state to another for a collaborative system.

Table 1. Transition by documents between states of collaborative systems

	S_1	S_2	...	S_i	...	S_n
S_1						
S_2						
...						
S_i				d_{ij}		
...						
S_n						

The collaborative system named *bank* covers the following states:

- S_1 – open;
- S_2 – money receive;
- S_3 – credit approve;

- S_4 – money discharge;
- S_5 – currency exchange;
- S_6 – closed.

Table 2 contains transaction between the six states of the *bank* collaborative system:

Table 2. Transaction between the six states of the *bank* collaborative system

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	void	yes	yes	yes	yes	yes
S_2	yes	void	yes	yes	yes	yes
S_3	no	yes	void	yes	yes	yes
S_4	no	yes	yes	void	yes	yes
S_5	no	yes	yes	yes	void	yes
S_6	yes	no	no	no	no	void

Table 3 contains transaction by command from one state to another for a collaborative system. Therefore, the change from S_i state to S_j state is done by C_{ij} command:

Table 3. Transaction by commands between collaborative systems states [2]

	S_1	S_2	...	S_i	...	S_n
S_1						
S_2						
...						
S_i				C_{ij}		
...						
S_n						

The transition from one state to another implies output delivery from the system. In a usual collaborative system changing from S_i state to S_j state is not possible for any i and j in $1..n$. The system changes from one state to the other but it doesn't have the capacity to go from each state in all the others. The situation is met for instance when a very simple collaborative system like a virtual mono-product store markets cement. The possible states of the system are: open, closed, supplying, sale. Table 4 contains the system transitions between the four states:

Table 4. Transition between states for a virtual mono-product store [2]

	Open	Closed	Supply	Sale
Open	void	yes	yes	yes
Closed	no	void	yes	no
Supply	yes	no	void	yes
Sale	no	yes	yes	void

The change from *open* state to *closed* is done through **close store** command, while the transition from *supply* to *open* is triggered by the **open store** command. Getting from *open* or *supply* states to *sales* is done through **merchandise** command.

Probabilities are assigned to such a collaborative system, like: probability to change from S_i state to S_j , probability to deliver x as output when changing from one state to

another. These probabilities are determined by a series of parameters like: amount of stock-in-trades, the deposit capacity of the store, working program, customers' portfolio or the amount of daily orders. The systems probability to transit from *open* state to *closed* is influenced by the following situations:

- is the end of the working program;
- the stock-in-trades is used-up, following change to *closed* state and then to *supply*;
- the clients' orders are not honored because of different reasons.

Probabilities of these transitions are used in determining the appearance frequencies of different states of the systems at a given moment. According to Table 4, the change from the state of *supply* to the same state is not possible. In certain situations – like a large order from one of the clients, the system will go from one supply to other in order to honor the customer's demand. But the probability of such situations is low, considering that one supply loads the whole deposit capacity of the store.

Let $P_{x_{ij}}$ be the probability that the collaborative system representing a virtual mono-product store provides output x when changing from S_i state to S_j state. $P_{x_{ij}}$ is determined by the relation:

$$P_{x_{ij}} = \frac{CF_x}{CP_x},$$

where:

$P_{x_{ij}}$ – probability that the collaborative system provides output x when changing from S_i state to S_j state;

CF_x – number of favorable cases to obtain output x when changing from S_i state to S_j state;

CP_x – number of possible cases to obtain output x when changing from S_i state to S_j state.

The bank collaborative system contains identities which generate messages like: demands for problem solution, open accounts, currency discharge, credit approval.

The identities are professional and in the job description they make certain operations. The person P_i operates n_i transactions: $O_{i1}, O_{i2}, \dots, O_{ini}$

Each transaction implies certain documents: $d_{i1}, d_{i2}, \dots, d_{ini}$

Each transaction has a solution: $s_{i1}, s_{i2}, \dots, s_{ini}$.

All of these are stored in a very large database.

The dynamics of the collaborative systems point to modifications in the quality, structure, functions, dimensions, procedures and standards of the systems. The dynamics are studied by mathematical analysis, forecasting the long term behavior of each system [4].

2. The database structure

In [5] a collaborative system is presented for very large datasets visualization using web services. Web services implement collaboration and visualization through internet of very large data sets.

An article from the very large database contains: name of the person, the operation, the time of demand reception, the time of solution delivery (the time between the two times is the solving period) and the solution given.

```
struct article
{
char *person;
char *demand;
int input_moment;
int output_moment;
char *product;
};
```

Person: *Johnson*

Demand: *Taxi-cab authorization*

Input moment: *10*

Output moment: *11*

Product: *Yes*

This data is recorded by:

- the solicitor who tells his demand; the moment of demand reception is stored along with the address where the solution has to arrive;
- the solution giver;
- the auditor who verifies the quality of system activity.

The system activity leads to generation of C collection with N elements, N very large ($N > 10^5$). The elements e_1, e_2, \dots, e_N have identical structure and conform with the unique description template. It contains the list of the field-attributes of each article, the order of appearance and the type of each attribute. The elements are updatable and the rate r of field update is known.

Let L be the number of locations where the N elements are placed and let $l_1, l_2, \dots, l_i, \dots, l_L$ be the number of elements belonging to the i location, so that $\sum_{i=1}^L l_i = N$. The local

collectivities are formed through determined aggregation criteria: geographical, structural, rank, etc. The local collectivities c_1, c_2, \dots, c_L are distinct physical databases.

Because of the large number of elements of any local collectivity c_i , the copy cost of elements is noticeable high. Therefore, in order to create the initial C collectivity with N elements a virtual concatenation of the L distinct databases is chosen. The central database integrates local collection by their description, functioning like a database concordance. Each record has a physical address and its own update data. In the moment when an update occurs, the data about it is stored in the local database c_i by modifying a single file which contains data about all the records in the collection. The concatenated database is updated by the automatic or manual copy of the L content files, when they are altered. The content files of the local databases can be database concordances of other smaller database of inferior hierarchic level.

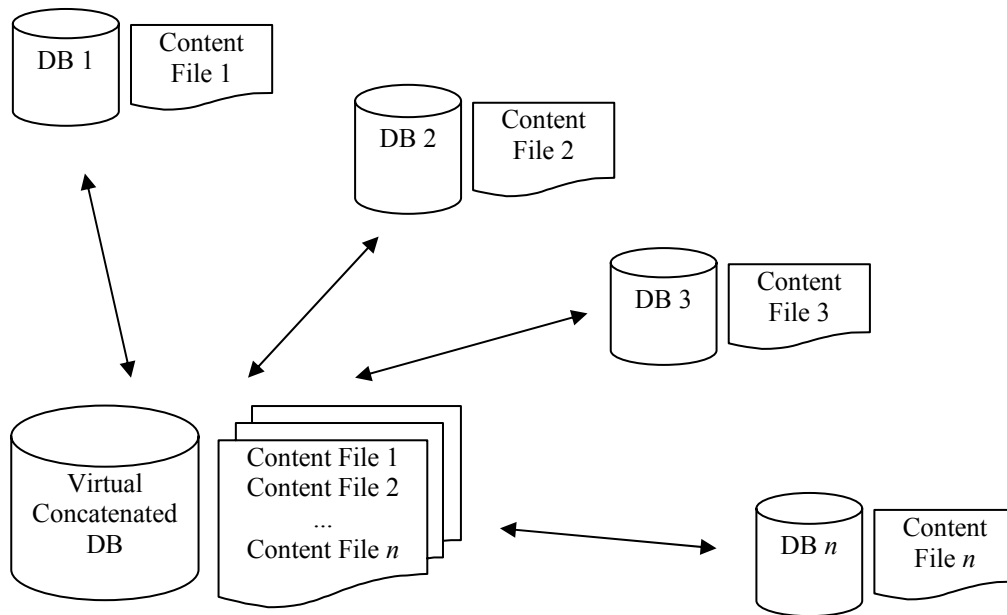


Figure 1. Virtual aggregation of DB

The daily data volume from the virtual database follows the relation:

$$VD_z = NP * \sum_{i=1}^{ND} NC_i, \text{ where:}$$

- NP – number of persons;
- ND – number of documents;
- NC_i – number of fields in the *i* document.

Data volume for a *k* days period is:

$$VD_{k\text{-days}} = k * VD_z, \text{ where:}$$

VD_z – daily data volume.

Let *DB* be a database whereat *N* stations have access for creating and updating records. In the $[t_i; t_{i+1}]$ time period there are P_i persons that operate updates on the database, $i=1, \dots, N$. The total number of persons that interacts with *DB* in the $[t_i; t_{i+1}]$ period is modeled by the relation:

$$NTP_j = \sum_{i=1}^N P_i$$

If each P_i makes k_i operations on *BD* and each k_i materializes in a new record in *DB*, then, the total number of records generated in the $[t_i; t_{i+1}]$ interval is:

$$NTI_j = \sum_{i=1}^N P_i k_i$$

Therefore, for any two intervals $[t_i; t_{i+1}]$ and $[t_{i+1}; t_{i+2}]$ the number of records is exactly determined by the addition of the above indicators. Through gradual aggregation the long time generated data is determined.

This daily or periodic data volume includes an image of all that is realized in the collaborative systems in the informational field.

The records in the database are done:

- online by the solicitor;
- online through documents management by the agents of the collaborative system.

The documents' processing means extracting the information from the data fields and converting as strings of bits, which can be stored in a database. The processing of forms is considered completed when all the information from documents has been extracted and saved within a database [4].

In [6] a mechanism for auto-organization of devices collection is presented, for processing documents in a collaborative system from a multi-agent perspective.

The input data quality determines the quality of the information that is given to the users or to the decisional authorities within the system. The erroneous data has to be minimized from the initial phase. The validation procedures are grouped in:

- traditional procedures which consist in visual examination of data in primary documents;
- automated data control and validation procedures, using validation software and automated data correction procedures; the errors are automatically identified and corrected without any human intervention [4].

The forms contain required and optional fields. The input data is consisted of: letter strings, dates, codes or numerical values. This data frames in the following categories:

- correct and complete; a form contains the fill-in fields C_1, C_2, \dots, C_n ; each field C_i has a value domain V_i ; the *correct* and *complete* state is achieved when all the fields $C_i, i=1..n$, belong to vocabularies $V_i, i=1..n$;
- correct but incomplete; all the fields $C_1, C_2, \dots, C_{i-1}, C_{i+1}, \dots, C_n$ belong to vocabularies $V_i, i=1..n$, but field C_i is missing; in this case, the application generates messages which indicate those fields which were not filled in;
- complete but incorrect; all the fields are filled in, but one or more values doesn't belong to the value-domain; in this case, the message generated specifies for each field what went wrong.

For each attempt of recording data, if the data is not completely correct and complete, a list of errors appear and the data is not sent to the database [7].

The application Collaborative Multicash Servicedesk records and processes phone demands taken by the analysts in a bank.

The fields that are selected or completed by the analyst are the following:

- name of the client which is selected from a predetermined list of Multicash users;
- name of the person which called;
- demand category, by selecting from a list with categories and associated codes;
- demand description for problem details;
- solving method by selecting the adequate option.

When a negative resolution is adopted to a credit demand, the justification is based on:

- incomplete documents in the credit;

- failing to fulfill the existing law requirements;
- failing to fulfill the technological restrictions: the wage is too small or assurance is inexistent.

Denials are coded for storing and for easy telling of the reasons of decision. Based on these codes reports and statistics are generated for classifying the denials on the type of negative responses.

3. Processing

Based on the very large datasets, many processing operations are realized in order to correlate and compute certain indicators.

Groups of persons are established and sorting takes place on the persons in the database leading to:

Name₁: t₁ articles;

Name₂: t₂ articles;

...

Name_k: t_k articles.

In the collaborative system there are *k* agents. From the Collaborative Multicash Servicedesk database certain data has been extracted about the number of demands, by clients recorded in November 2009, presented in Table 5:

Table 5. Number of demands by clients

Client	Number of demands
Client Name 1	200
Client Name 2	180
Client Name 3	220
Client Name 4	100
Client Name 5	300

Analyzing the number of demands by clients its determined the number of distinct clients that recorded demands in a specific period.

In the present, is seeking the increase of services quality offered by collaborative banking systems, by introducing intelligent agents to help increase performance of these systems.

The intelligent agent means an entity performing certain operations independently on behalf of a third party. The agents have a number of attributes, their main attribute is autonomy. For an agent to be called intelligent, its autonomous nature must be flexible, meaning that:

- perceive the working environment and the appropriate response to the changes occurred;
- decide to action also in situations of not environment amending;
- ability to interact with other agents or even with the human agent, both to achieve the designed goals and to facilitate the work of other employees.

Characteristics of an agent are the followings:

- mobility, which is defined as the ability to move in an electronic network;
- veracity, which implies that an agent is unable to provide false information;

- goodwill, which means that an agent does not have conflicting goals;
- rationality, which means that an agent acts to achieve the purpose.

An agent is characterized by an architecture and a program. The agent program is a function that matches the perceptions which the agent receives from the environment and its actions. This program must be compatible with the agent architecture. The architecture made the interface between the perception given by sensors and the program, run the program and ensure that all actions chosen are made as they are generated. The environment where an agent act has many facets, being fully or partially observable, deterministic or stochastic, static or dynamic, discrete or continuous, monoagent or multiagent.

The sort is done by messages or documents received and results:

Docum₁: k₁ appearances;

Docum₂: k₂ appearances;

...

Docum_h: k_h appearances.

It follows that in the system are *h* types of documents.

They are sorted by resolution and appear:

Yes: k₁ appearances;

Positive opinion: k₂ appearances;

...

Rejected: k_r appearances.

It follows that in the collaborative systems are *r* types of resolutions.

In the same manner are analyzed the evidence answers.

Incomplete documents: x₁ appearances;

Violation of legal provisions: x₂ appearances;

Technological incompatibilities: x₃ appearances.

Table 6 presents a report from the database of Collaborative Multicash Servicedesk application, regarding the categories of requests and their frequency in the month of November 2009:

Table 6. Number of requests by category

Category	Number of requests
Add new accounts in the client application	26
Add new users in the client application	14
Other requests	132
User blocked on the communication	41
User blocked at logon	20
Communication initiated	54
Confirm account balance	71
Deactivate payments file	1
Error on starting the application	5
Signature error	46
Error on see statements	20
Statements export	1
Generate electronic signature	20
Index corrupted in database tables	4

Category	Number of requests
Training on using the application	14
Training on see rejected payments	4
Delivery account statements	7
Delivery file with bank codes	12
Delivery files for distributed signature	8
Change communication channel	1
Change number of approvals / amount limits	1
Change name / address of payer	1
Move the application on another computer	13
Please repeat job with AC29	9
Reinstalling the application	7
Setting print parameters	5
Setting communication sessions	1
Training of branches for completing annexes	10
Transmission interrupted	36
Check payments status	162

On the basis of processing performed on the data sets, sorting is made by a single feature and are determined the metrics:

- average number of documents per person, *NDP*:

$$NDP = \frac{NTD}{NP}, \text{ where:}$$

NTD – total number of documents in the system;

NP – total number of persons.

- average number of refusals to 100 requests, *NR*:

$$NR = \frac{NTR}{NS} * 100, \text{ where:}$$

NTR – total number of refusals in the system;

NS – total number of requests.

Other statistics are performed in order to be used in the justice to determine the correlations between documents, customers, requests, and resolutions.

4. Combined analysis

Data sets are identified and is performed a combined analysis to determine certain statistics. The combined analysis involves correlations between data sets, for the calculation of quality indicators.

There are considered *M* large databases *BD₁, BD₂... BD_M*, made with the same informatics application, reflecting data which characterize *M* collectivities ordered in disjoint territorial areas. It builds an informatics application for realize the virtual database *BDV* by an operation of concatenation of the basic data extracted from databases *BD₁, BD₂... BD_M*.

The selection of records from the virtual database *BDV* requires to:

- across the data set of essential results in the concatenation process;
- identify the BD_i databases containing records associated to the selected essential data;
- take information from real databases for processing selected records;
- carrying out processing.

For the analysis Person – Operations, are identified the types of operations made by a person:

Popescu: settle rents documents, settle taxi permits, resolve global tax.

Ionescu: settle construction permits.

Is determined the load degree of each agent in the system and is made a redistribution of operations so that do not exist a situation in which an agent is overloaded and another do not have enough operations which fill the working time.

From the combined analysis Analyst – Category of requests, on the basis of records from the Collaborative Multicash Servicedesk application, results that the analyst *Mihai Iancu* solved requests from the categories *Add new accounts in the client application, User blocked on the communication, Generate electronic signature, Change communication channel*, and the analyst *Marian Neagu* solved requests from the categories *Add new users in the client application, Training on see rejected payments, Move the application on another computer*. Taking into account the number of requests recorded on each category, it follows that the analyst *Mihai Iancu* has been overloaded.

For the analysis Person – Resolutions, there are evaluated the types of resolutions adopted and their frequencies of occurrence:

Popescu: resolution YES at the rate of $x\%$, NO at the rate of $y\%$.

Ionescu: resolution YES at the rate of $z\%$, NO at the rate of $w\%$.

If $x > z$, then Popescu gave more positive resolution than Ionescu. If $x > y$, then Popescu gave more positive than negative resolution. If $z > w$, then Ionescu gave more positive than negative resolution.

By generalization, being considered the data sets D_1, D_2, \dots, D_n , correlations are established between any of D_i and D_j , where $i, j = 1..n$, with $i \neq j$. For each combined analysis $D_i - D_j$ the types of correlations are analyzed and are calculated quantitative and qualitative indicators.

Indicators for the case presented above are:

- the quantitative indicator comparing the number of resolutions adopted by the two entities:

$$I_{D_i/D_j} = \frac{N_{D_i}}{N_{D_j}}, \text{ where:}$$

N_{D_i} – the total number of resolutions adopted by D_i ;

N_{D_j} – the total number of resolutions adopted by D_j ;

- the qualitative indicators comparing values between the two resolutions adopted:

$$I_{D_i} = \frac{x}{y}; \quad I_{D_j} = \frac{z}{w};$$

$$I_{x/z} = \frac{x}{z}; \quad I_{y/w} = \frac{y}{w};$$

From the calculation of the quantitative and qualitative indicators result the current status of collaborative systems and the elements requiring replacement or improvement.

5. Construction of procedures for online problem solving

The document is followed from entry until it will be solved in order to see what is doing. There are identified the routines and activities that are repetitive. The objective of online problem solving procedures is to automate the repetitive activities in order to increase the response time and to resolve the requests.

The collaborative systems have: structure, purpose, flows, resources, routines.

For the real organization that works is built its entry into the database, from which are taken rules, using large volumes of data on which are made selections, sorting, regrouping, extractions.

From reality is issued the collaborative system model, identifying:

- the number of real states;
- the real list of states;
- types of resources;
- the real list of resources;
- number of activities types;
- list of activities;
- durations or differences between final and initial moments.

It follows the real image of the system, from which are obtained the basic parameters of the model, average durations, limitations of resources, on which the model is built.

The collaborative system has associated a neural network giving automatic solutions.

The neuronal network deliver output data and the input data are taken from databases.

Having start and final moments available, the durations are determined and, by their size, frequencies are calculated.

There are determined the probabilities that durations will be equal to a given value, $P(\text{duration} = X) = Y$, or the durations to be less that a given value, $P(\text{duration} < Z) = W$.

A customer arrives with a request, enter to the portal, is called the neuronal network, resulting the proposed solution which is provided to the customer and he decides. This workflow is shown in Figure 2:

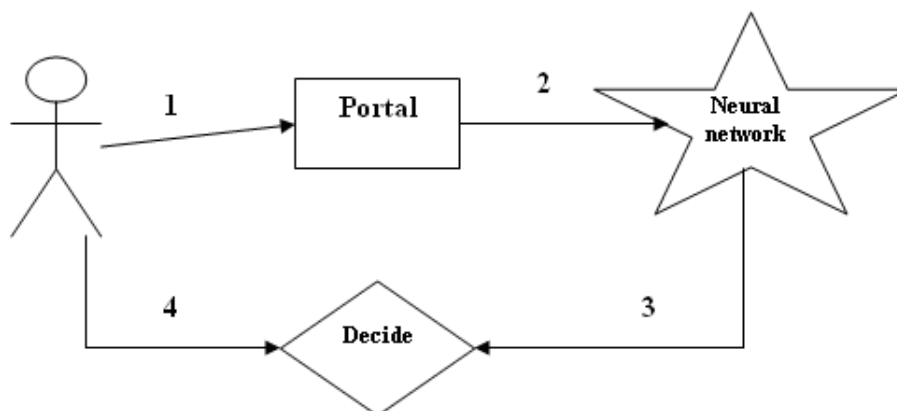
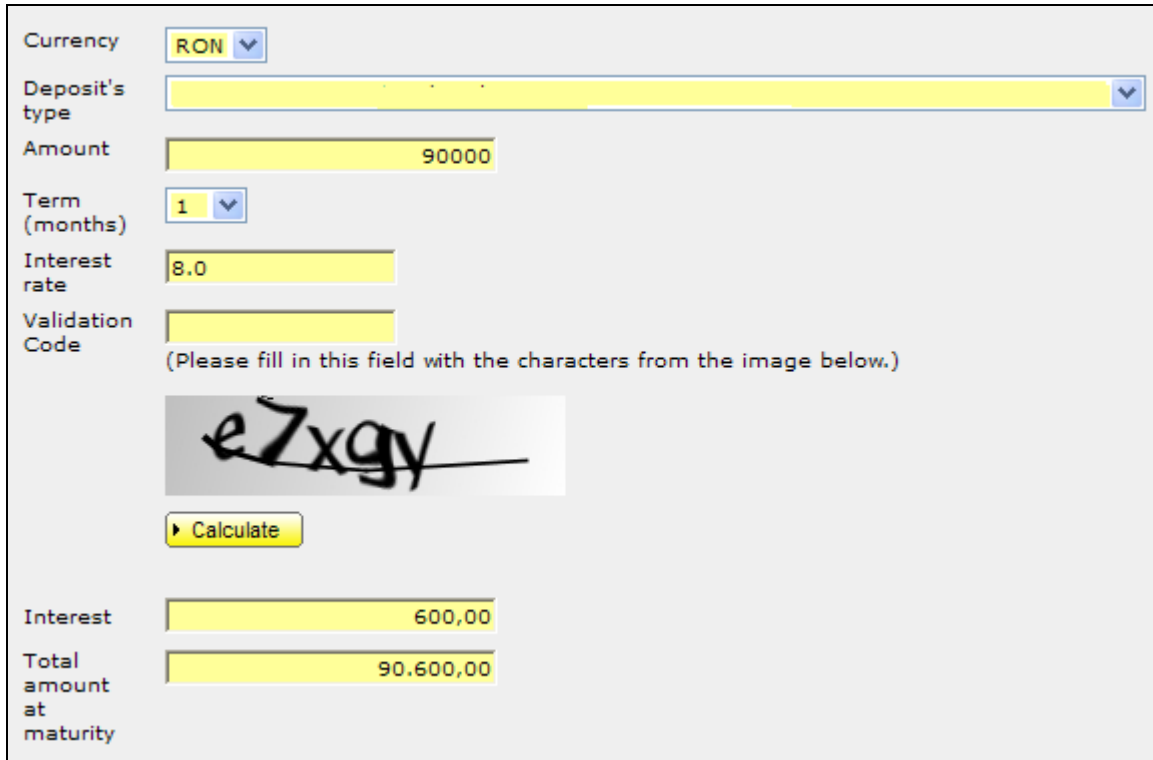


Figure 2. Workflow in a collaborative system

In a collaborative banking system, the customer interaction with the bank is done in several ways, some of which requiring a minimal effort from the client. On the banks websites are a series of applications and simulators to calculate interest rates on loans or deposit. Figure 3 shows an example of simulator for the calculation of receivable interest related to a term deposit:



Currency:

Deposit's type:

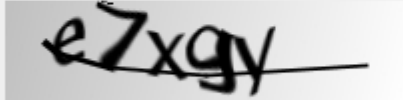
Amount:

Term (months):

Interest rate:

Validation Code:

(Please fill in this field with the characters from the image below.)



Interest:

Total amount at maturity:

Figure 3. Deposits calculator (<http://www.raiffeisen.ro>)

Another example of simulator is an application for credit related calculation. The customer enters the bank website, select the type of loan, grant period, currency and interest type and enter the amount required.

The application will display to the user the monthly rate and amount of fees. These calculations are carried out also in the bank agencies, but this involves the customer go to the bank.

6. Conclusions

Collaboration is better appropriated if is based on simple rules, leaving the agents to fulfill their interests within their societies.

Increasing the volume of information and improving the software products for exploit it have led to a new quality of data usage by analysis that reveal to the organization's management information difficult or impossible to obtain otherwise. In this way are obtained information on customer preferences, their profile or distribution.

A collaborative system is characterized by a diversity of states, its transition from one state to another being accomplished through a document or command. The set of collaborative system states is finite, with an initial state and a final one. The system

throughput between the initial state, the intermediate states and the final state, carrying out a cycle when it complete the transition from the final state to the initial state.

References

1. Dobrican, O. **An Example of Collaborative System**, International Workshop Collaborative Support Systems in Business and Education, Risoprint, Cluj-Napoca, October 2005, pp. 48
2. Ivan, I., Boja, C. and Ciurea, C. **Collaborative Systems Metrics**, Bucharest: ASE Publishing House, 2007
3. Ivan, I. and Ciurea, C. **Quality Characteristics of Collaborative Systems**, International Conference on Advances in Computer-Human Interaction, ACHI 2009, pp. 164-168, 2009, Second International Conferences on Advances in Computer-Human Interactions, 2009
4. Ivan, I. and Ciurea, C. **Using Very Large Volume Data Sets for Collaborative Systems Study**, Informatica Economica Journal, Vol. 13, No. 1, 2009
5. Zhao, Y., Hu, C., Huang, Y. and Ma, D. **Collaborative Visualization of Large Scale Datasets Using Web Services**, Internet and Web Applications and Services, 2007, ICIW '07, Second International Conference on, pp.62-62, 13-19 May 2007
6. Gnanasambandam, N., Sharma, N., Kumara, S. R. and Hua, L. **Collaborative Self-Organization by Devices Providing Document Services - A Multi-Agent Perspective**, Autonomic Computing, 2006, ICAC '06, IEEE International Conference on, pp. 305-308, 13-16 June 2006
7. Ivan, I., Ciurea, C. and Milodin, D. **Validarea datelor de intrare în aplicațiile informatice orientate spre cetatean**, Revista Romana de Informatica si Automatica, Vol. 18, Nr. 4, 2008

¹ Acknowledgements

This article is a result of the project POSDRU/6/1.5/S/11 „Doctoral Program and PhD Students in the education research and innovation triangle”. This project is co funded by European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies, project no. 7832, Doctoral Program and PhD Students in the education research and innovation triangle, DOC-ECI.

²Ion IVAN has graduated the Faculty of Economic Computation and Economic Cybernetics in 1970. He holds a PhD diploma in Economics from 1978 and he had gone through all didactic positions since 1970 when he joined the staff of the Bucharest Academy of Economic Studies, teaching assistant in 1970, senior lecturer in 1978, assistant professor in 1991 and full professor in 1993.

Currently he is full Professor of Economic Informatics within the Department of Computer Science in Economics at Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies. He is the author of more than 25 books and over 75 journal articles in the field of software quality management, software metrics and informatics audit.

His work focuses on the analysis of quality of software applications. He is currently studying software quality management and audit, project management of IT&C projects. He received numerous diplomas for his research activity achievements.

For his entire activity, the National University Research Council granted him in 2005 with the national diploma, Opera Omnia. He has received multiple grants for research, documentation and exchange of experience, conferences and congresses at numerous universities from Greece, Ireland, Germany, France, Italy, Sweden, Norway, United States, Holland, Australia, China and Japan.

He is distinguished member of the scientific board for the magazines and journals like:

- Economic Informatics
- Economic Computation and Economic Cybernetics Studies and Research
- Romanian Journal of Statistics

He has participated in the scientific committee of more than 20 Conferences on Informatics and he has coordinated the appearance of 3 proceedings volumes for International Conferences. From 1994 he is PhD coordinator in the field of Economic Informatics.

He has coordinated as a director more than 15 research projects that have been financed from national and international research programs. He was member in a TEMPUS project as local coordinator and also as contractor in an EPROM project. Also, he was expert assessor in many research programmes like RELANSIN, INFOSOC, CALIST and CEEEX. He holds many awards and diplomas in research activity.

His main interest fields are: software metrics, optimization of informatics applications, developments and assessment of the text entities, efficiency implementation analysis of the ethical codes in informatics field, software quality management, data quality management and so forth.

List of Main Publications (2006 – 2009)

- Ion IVAN, Marius POPA and Paul POCATILU (coordinators) – **Structuri de date**, ASE Printing House, Bucharest, 2008, vol. I Tipologii de structuri de date; vol. II Managementul structurilor de date
- Ion IVAN and Catalin BOJA **Practica optimizarii aplicatiilor informatice**, ASE Printing House, Bucharest, 2007
- Ion IVAN, Catalin BOJA and Cristian CIUREA **Metricile sistemelor colaborative**, ASE Printing House, Bucharest, 2007
- Ion IVAN, Gheorghe NOSCA, Sergiu CAPISIZU and Marius POPA **Managementul calitatii aplicatiilor informatice**, Bucharest: ASE Printing House, 2006
- Ion IVAN, Traian BADICA and Marius POPA **Procese de agregare software**, Studii si Cercetari de Calcul Economic si Cibernetica Economica, vol. 42, no. 1, 2008, pp. 69 – 85
- Ion IVAN, Eugen DUMITRASCU and Marius POPA **Evaluating the Effects of the Optimization on the Quality of Distributed Applications**, Economic Computation and Economic Cybernetics Studies and Research, vol. 40, no. 3-4, 2006, pp. 73 – 85
- Ion IVAN, Cătălin BOJA, Marius VOCHIN, Iulian NITESCU, Cristian TOMA and Marius POPA **Using Genetic Algorithms in Software Optimization**, Proceedings of the 6th WSEAS International Conference on Telecommunications and Informatics, Dallas, Texas, USA, March 22 – 24, 2007, pp. 36 – 41
- Ion IVAN, Marius POPA, Catalin BOJA, Cristian TOMA and Dragos ANASTASIU **Software for Structured Text Entities Dependency Graph Building**, Proceedings of the 2007 WSEAS International Conference on Computer Engineering and Applications, Gold Coast, Queensland, Australia, January 17 – 19, 2007, pp. 224 – 230
- Ion IVAN, Cristian TOMA, Marius POPA and Catalin BOJA **Secure Architecture for the Digital Rights Management of the M-Content**, Proceedings of the 5th WSEAS International Conference on Information Security and Privacy, Venice, Italy, November 20 – 22, 2006, pp. 196 – 201

³ Cristian CIUREA has a background in computer science and is interested in collaborative systems related issues. He has graduated the Faculty of Economic Cybernetics, Statistics and Informatics from the Bucharest Academy of Economic Studies in 2007. He is currently conducting doctoral research in Economic Informatics at the Academy of Economic Studies. Other fields of interest include software metrics, data structures, object oriented programming in C++ and windows applications programming in C#.

⁴ Sorin PAVEL has graduated the Faculty of Economic Cybernetics, Statistics and Informatics from the Bucharest Academy of Economic Studies in 2008. He is currently following Master's in Software Project Management and the Doctoral School in Economic Informatics, both at the Academy of Economic Studies.

⁵ Codification of references:

[1]	Dobrican, O. An Example of Collaborative System , International Workshop Collaborative Support Systems in Business and Education, Risoprint, Cluj-Napoca, October 2005, pp. 48
[2]	Ivan, I., Boja, C. and Ciurea, C. Collaborative Systems Metrics , Bucharest: ASE Publishing House, 2007
[3]	Ivan, I. and Ciurea, C. Quality Characteristics of Collaborative Systems , International Conference on Advances in Computer-Human Interaction, ACHI 2009, pp. 164-168, 2009, Second International Conferences on Advances in Computer-Human Interactions, 2009
[4]	Ivan, I. and Ciurea, C. Using Very Large Volume Data Sets for Collaborative Systems Study , Informatica Economica Journal, Vol. 13, No. 1, 2009
[5]	Zhao, Y., Hu, C., Huang, Y. and Ma, D. Collaborative Visualization of Large Scale Datasets Using Web Services , Internet and Web Applications and Services, 2007, ICIW '07, Second International Conference on, pp.62-62, 13-19 May 2007
[6]	Gnanasambandam, N., Sharma, N., Kumara, S. R. and Hua, L. Collaborative Self-Organization by Devices Providing Document Services - A Multi-Agent Perspective , Autonomic Computing, 2006, ICAC '06, IEEE International Conference on, pp. 305-308, 13-16 June 2006
[7]	Ivan, I., Ciurea, C. and Milodin, D. Validarea datelor de intrare în aplicațiile informatice orientate spre cetatean , Revista Romana de Informatica si Automatica, Vol. 18, Nr. 4, 2008

INTERRELATIONSHIPS OF ORGANIZATION SIZE AND INFORMATION AND COMMUNICATION TECHNOLOGY ADOPTION

Mladen CUDANOV¹

MTS, Teaching Assistant, Faculty of Organizational Sciences,
University of Belgrade, Serbia

E-mail: mladenc@fon.rs, **Web-page:** <http://www.linkedin.com/in/mladencudanov>



Ondrej JASKO²

PhD, Associate Professor, Faculty of Organizational Sciences,
University of Belgrade, Serbia

E-mail: jasko@fon.rs,

Web-page: <http://www.fon.rs/ofakultetu/nastavnici/CV/JaskoOndrej/index.html>



Gheorghe SAVOIU³

PhD, Associate Professor, Faculty of Economics,
University of Pitesti, Romania

E-mail: gsavoiu@yahoo.com,

Web-page: <http://ro.linkedin.com/pub/gheorghe-savoiu/16/337/b27>



Abstract: *This paper aims to describe interrelationships between size of the organization and adoption of information and communication technologies (ICT). We hypothesize that size of the organization is interrelated with ICT usage. By analyzing the data from 68 organizations we have classified to micro, small, medium-sized and large enterprises and calculated composite index of ICT adoption in each organization. Afterwards we have analyzed correlations between composite index of ICT adoption and size of the organization. Our results show that large enterprises which have potential to utilize ICT have highest values of composite index of ICT adoption, indicating high ICT usage. Theory considered in the discussion implies that ICT diminishes size of the organization, which complies with our findings because medium enterprises keep high values of composite index of ICT adoption. Small organizations, at least in transitional countries, in average do not show high level of ICT use, but especially in smallest, micro organizations extreme examples both of high and low ICT use indicated by high standard deviation values can be found. That could be explained by greater flexibility and orientation of small enterprises to new technologies, but also lack of resources or interest and implies that in small and micro companies ICT appliance is more dependent on other organizational factors than on size. Our conclusion is that ICT has the potential to diminish size of the company, but that still in average large and medium companies are leaders of ICT use in spite of extreme examples of good practice in small companies.*

Key words: Organization size; ICT adoption; Composite index of ICT adoption; Organizational factors

1. Introduction

This article aims to describe interrelationships between size of the organization and applications of information and communication technologies (ICT). Size can be observed as important factor of implementation of ICT. Small and medium enterprises can be important innovation factor in economy, and Van Dijk et al. (1997) observe them to be more open to applying innovations⁴. They also have better use of innovations developed in large organizations and scientific institutes (Audretsch & Vivarelli 1996).⁵ Common attitude is that small and medium enterprises are more open to ICT than large enterprises, and are ready to form alliances when size does not permit using technology advancements (Narula 2001)⁶. Increased implementation of ICT can be consequence of such alliances of small and medium enterprises.⁷ Small and medium enterprises can lower the costs of ICT appliance by using open source alternatives, and even in developing and underdeveloped countries illegal copies of expensive software (Van Belle&Ellis 2009).⁸

On the other hand, large enterprises have its own factors that facilitate implementation of information and communication technologies, and most important can be critical financial mass that enables access to expensive technologies and management requirements that imply larger demand for ICT use. The potential of large enterprises to use ICT is described in detail by many authors, such as Cisler(2005)⁹, Marmaridis(2005)¹⁰, Trimi(2005)¹¹ or Kramer et al (2007)¹². That assumption is justified, but it is important to observe that it influences only potential for use, but not always the real use of information and communication technologies.

The idea that size of the organization can influence implementation of information and communication technologies comes from theoretical view that the size of the enterprise is important organizational factor (Blau and Schoenherr 1971¹³; Mintzberg 1980¹⁴). Other organizational features, such as decentralization, work division, departmentalization and coordination are related to information and communication technologies (Bloomfield & Coombs 1992¹⁵; Čudanov 2007)¹⁶, so it is rational to create hypothesis that organizational size can be related to ICT.

2. Methods

Most common enterprise size indicators are total number of employees, total assets and enterprise income. For the purpose of this article we accepted all three indicators and started from classification of small, medium sized and large enterprises, according to Serbian law that uses:

- Average number of employees
- Total income
- Assets value stated in financial report in last business year¹⁷

This division has been widened by "Micro" category, that have two out of three conditions: less than 10 employees, total income less than 50 milion od RSD (circa 500,000 EUR) and assets under 10 milion of RSD (Circa 100,000EUR). Small enterprises are those that comply with two out of three conditions: number of employees is between 10 and 50, income between 50 and 225 milion of RSD (circa 500,000-2,25milion EUR) and assets between 10 and 90 milion of RSD (circa 100,000-900,00EUR). Medium enterprises are those that comply with two out of three conditions: between 50 and 250 employees, income

between 225 and 900 million of RSD (circa 2,25-9 million EUR) and assets between 90 and 450 million of RSD (circa 900,000-4,500,000EUR). Large enterprises are those that comply with at least two of three conditions: more than 250 of employees, income above 900 million of RSD (circa 9 million EUR) and assets value above 450 million of RSD (circa 4,5 million EUR).

This division resulted in following distribution of enterprises, with organizations sorted into industries:

Table 1. Structure of the sample of organizations used in analysis

Size/industry	ICT	Production	Commerce	Services	Total
Micro	6	7	3	5	21
Small	3	8	2	4	17
Medium	2	7	8	3	20
Large		7	1	2	10
Total	11	29	14	14	68

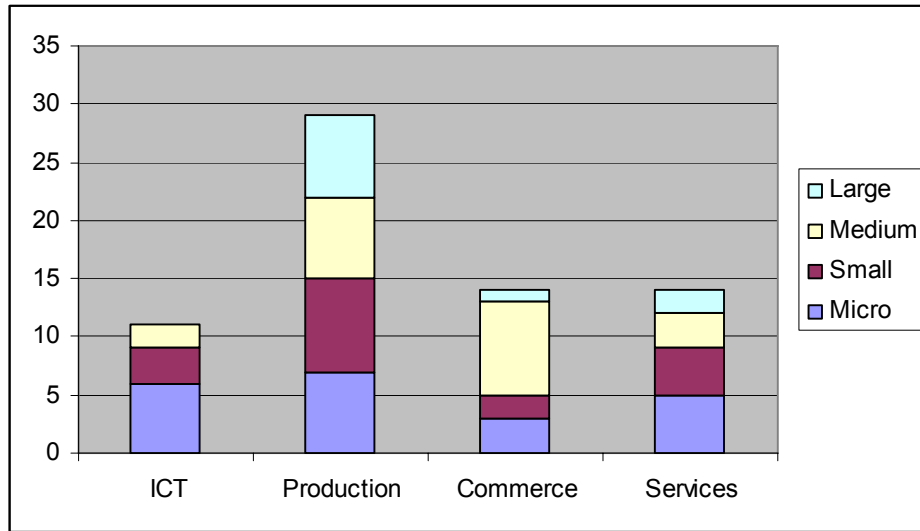


Figure 1. Structure of the sample used for analysis

As a second factor in analysis, we have used Composite index of ICT adoption. As an alternative to subjective assesment several indicators connected to the use of ICT in the company were selected from quantitative data gathered. Selected indicators, different in nature, were combined and a composite index of ICT adoption (abbreviated CICT in tables) was formed. This formula has already been used to asses implementation of information and communication technologies in the enterprise (Čudanov et al. 2009)¹⁸.

Formula of the composite index is presented in following:

$$CICT = \frac{NoC}{NoE} + \frac{NoCC}{NoE} + \sum_{i=1}^{i=8} Cf_i + \prod_{i=1}^{i=8} \left(\sqrt{\frac{NoCC}{NoE}} + Cf_i \right) + CDB + DBA$$

Where mentioned

factors mean:

CICT	=	Composite index of ICT adoption in company;
NoC	=	Number of computers in the company;
NoE	=	Number of employees in the company;
NoCC	=	Number of computers connected to internal network in the company;
Cf _i	=	Coverage of enterprise function by ICT, where for different values of <i>i</i> functions are: 1 - human resources, 2 – accountancy, 3 – financial, , 4 – technical, 5 – commercial, 6 – administrative, 7 – legal, 8 – protection; Coverage of business function was estimated by IT staff, functional staff and top managements as percentage of usual job in that function supported by ICT existing in the organization
CDB	=	Existence of integrated company database (0=no, 1=yes)
DBA	=	Database administrator present (0=no, 1=yes)

After that, average values and variance has been analyzed to observe differences between groups. Additional descriptive statistics have been analyzed, also. Further sample has been analyzed by ANOVA method of variance analysis, tolerant for small variations of sample from normal distribution. In this case those variations were result of larger number of groups in total sample.

3. Results

First group of enterprises that has been distinguished by the value of composite index of ICT adoption consists of 17 small enterprises with very low value of that index - 4,12. This value can be explained by lack of resources needed for coverage of small business by information and communication technologies. Business of small enterprise is significantly more complex than in micro enterprises, although sole number of employees does not lead to such conclusion. It must be valued that complexity of such systems is more exponentially than linearly rises in accordance with rise in number of employees. Companies with up to 50 employees are in average much more complex systems compared to companies with up to 10 employees, because such companies represent heterogeneous business systems, that most often cover all basic business functions, like finance, production (or services) and sales, and such complexity demands significant support of information and communication technologies compared to micro enterprises. Low values of standard deviation give us good confidence about small variations of low value of composite index of ICT adoption.

Micro enterprises have significantly larger value of composite index of ICT adoption. Its value is 20,86, but high standard deviation of 41,95 implies that appliance of information and communication technologies has much more oscillations in such systems. It is assumed that such oscillation is due to large differentiation in micro business orientation toward ICT, that is largely dependable on industry and attitude of the owner towards ICT. High value of that index originates in simplicity of coverage of micro business with information and communication technologies. In accordance to division presented in methods section, micro enterprises are often organizations or even single entrepreneurs that perform single business function, without structurally differentiated departments that increase complexity of the system. Such business can be supported by simple patterns that can be created in mass accessible software. In simple business systems large number of administrative tasks can be performed sut ICT support from MS Office®¹⁹ applications or similar software alternatives, among which some are distributed free of charge as "open source" software. So it is reasonable to say that Cfi, or "Coverage of enterprise function by

ICT" can be fully or mostly covered in most of the human resources, accountancy, financial, technical, commercial, administrative, legal, protection functions of micro business. Further study of such potentials can improve communication and collaboration, within organization or outside its boundaries.

Third group of enterprises is consisted by medium and large enterprises. Those enterprises have high and very similar values of composite index of ICT (respectively 34,28 and 34, 65). Standard deviation is very high, respectively 79,30 and 86,62. Such values imply relatively better position that medium and large enterprises have in appliance of ICT, but such results should be observed with reserve, because sample is consisted of transitional companies, and some conclusions might be different in developed economies, where small and medium enterprises have large tradition and stronger resource support than in transitional countries. In context of transitional countries, however, like Serbia, Bulgaria or Romania, such results are largely confirmed by business practice. Medium and large enterprises have enough resources to try to achieve competitive advantage, and often include ICT as a tool for such goal, while small and micro enterprises resort to comparative advantage, local scope and small market niches, that do not demand implementation of ICT. Following table represents results of descriptive statistical analysis of our sample.

Table 2. Descriptive statistical analysis for composite index of ICT adoption in selected groups

	N	Average	Std. deviation	Std. error	95% interval of confidence for average value		Min.	Max.
					low bound.	high bound.		
Micro	20	20,86	41,95321	9,38	1,22	40,49	,14	173,31
Small	17	4,12	3,25391	,79	2,44	5,79	,24	10,72
Medium	20	34,28	79,30483	17,73	-2,83	71,40	1,36	322,15
Large	10	34,65	86,62009	27,39	-27,31	96,62	1,39	280,43
Total	67	22,68	59,12052	7,22	8,26	37,10	,14	322,15

Further analysis by ANOVA is did not confirm that observed differences are statistically significant, but non-parametric test have indicated statistical significance. That directs us toward possible repetition of such experiment on larger set of enterprises. Mixed results have also been observed by Lee and Xia (2006)²⁰, and they explained such results by unclear definition of factors of IT innovation, that is only part of ICT implementation in organization

4. Discussion

Results of this study indicate that size of the organization and implementation of ICT in that organization are connected. Usual approach regarded average number of employees as independent variable, while the other indicators are regarded as dependent. The situation which changed by automation, reengineering and „dot.com“ boom caused the change of such attitude, and excessive agitation of influence that information and communication technologies have on the size of the enterprise, but the crash of dot.com concept that among other propagated flat and small organizations partly moderated that

attitude. However, positive examples of companies which survived dot.com crash indicate the change of traditional relations of these values.

Information and communication technologies appliance and change in company size can be observed in the light of other global economic shifts. Shifting the focus of from the manufacturing to service economy caused appearance of great number of new trends, like making values in services is often possible with less human work than in manufacturing. Many enterprises that do business on Internet are totally based on information and communication technologies and have unusually small number of employees. Virtual enterprise concept enabled by ICT change classical definition of organization, like there is no building with offices in which these workers work. A few people can maintain the web site at home or sit in a rented business premise, achieving business results which are in clusters with enterprises which have several tenths times more employees. Influence of information technology to the size of enterprise is realized indirectly through better interior communication, bigger inter-organizational cooperation and more possibilities of outsourcing.

Nevertheless, our consulting experience concluded that there is no significant long term reduction in number of required employees in particular company as a result of information and communication technologies adoption. Osterman (1986) believes that information technologies have tendency to increase number of required employees for a period of several years after it's introduction.²¹ In such case information and communication technologies use potential of increased productivity not to reduce the number of required workers but to cope with increased amount of work. In some cases information and communication technologies complement rather than replace work force, in particular white collar work force.

Theoretical approaches that connect information technology with all indicators of enterprise size date from the period of the first half of 1970's of last century, before boom of Internet and information technologies utilization. Even in 1973 Arrow regarded market and enterprises like entities which at the first place analyze information,²² what Galbraith (1977) clearly confirmed in his studies²³. This stand point implies that information and communication technologies must have important influence to all parameters of the enterprise. It was noticed the tendency that size of the enterprise shown in the number of employees significantly decrease from 1970's of last century, while until the 1970's, this trend was opposite (Piore & Sabel, 1984)²⁴. It is interesting that this trend linearly is not followed by other indicators of growth, especially total revenue of enterprise. A lot of studies connected ICT related factors with organizational performance (Weil 1994;²⁵ Wilcocks 1999²⁶; Čudanov et al. 2007²⁷). It was announced by the American bureau for job statistic that from the period of 1980 until 1986, enterprises with less than 100 employees created more than six million new jobs, while the enterprises with more than 1000 employees dismissed more than one and the half million workers. From 1980, many studies which have been published examine correlation between the facts that, in global, enterprise size is considerably decreasing, but the enterprise profit share invested into information technologies is considerably increasing.

Confirmation of such trend can be found in vivid example of Google inc. which was founded in 1997. After eight years, specific philosophy of doing business lead the company from total number of employees of 2 to 3.021 employee in December 2004 (which was sudden increase compared with 2.668 in September 2004), with total assets of

3,313,351,000USD and total revenue of 1,032,000,000USD. In 2008, company earned 21,795,550,000USD of total revenue, owned 31,767,575,000USD of total assets²⁸ but increased number of employees only on 19,665. The reduction of the size of organization influenced by information and communication technology is trend which is also noticed in the companies not directly linked to the e-business or Internet. Development of information and communication technologies automated many processes. The reduction of participation of human labor increased productivity and ranked ICT among the strategic advances of the company. These enterprises have less employees compared to the industrial average. Re-engineering of many processes in classic branches of doing business enables finishing more work with less people. ERP systems, automatic systems of purchase or some instruments of e-business additionally automate supporting systems, further reducing need for administrative and supportive staff. Information and communication technologies are also one of the factors which enabled virtual and network organization structures.

Major influence paths of ICT on size of the enterprise includes outsourcing and automation. First is largely connected to distribution of added value among several smaller entities, while second aims at increasing efficiency and decreasing input of human work in processes. Potentials of the information and communication technologies improve the practice of the outsourcing of some activities. Analyzing the costs of transactions described by Williamson (1979)²⁹ enables us to determine effective limits of the company, and the general trend of their reduction is explained through the factors of reducing the expenses of the external coordination, which is in fact influenced by ICT. On the other hand, information systems are usually made with the purpose of improving the internal, not the external coordination, and improving the internal coordination is one of the factors that should result in lower costs of these activities, and in the larger companies, which rather decide on production than on purchase of goods and favors. Reducing the costs of the external coordination is easily correlated with the decrease of an average size of the company. Even though the ICT cannot eliminate opportune behavior, they can diminish the problems created by such opportunism, and, primarily, can decrease the rational limitations that participants in the market have. Reducing the costs of activities related to gathering valid information and administrative activities that are affected by coordination with external suppliers, informational technologies cut down the costs of the external transactions.

If the products or services are specific, despite of the positive influence of the information technology, at some limit external coordination becomes inefficient and leads to the practice of expanding the effective limits of the company. But ICT can reduce the expenses of the external coordination through the change in the specifications of the goods that are objects of the trade. If the techniques that are enabled by the development of the information and communication technologies are used, costs for specifications of the goods and external coordination are diminished. Significant examples for this can be found in the car manufacturing industry, which was, for a longer period of time, in the first plan of the automation. When Nissan, for example, built their new factory in Sunderland, in north England, they invited their biggest suppliers to build their own factories in the circle around Nissan's factory area and in that way become a part of Nissan's production control system. The goal that illustrates trend of smaller enterprise by outsourcing was that suppliers deliver their parts directly to the production line, in order to save on the storage space and handling expenses. Not only that information system imposed reduced size of one large car

manufacturing entity to web of interconnected companies, but also size of those companies was reduced due to new collaboration technologies that were not so work intensive.

As global trend, compensation of the reduction of the costs of internal coordination could be explained through the process of globalization, which is also partly backed by modern information and communication technologies. Even though the costs of finding the supplier, negotiating, contracting and payment can be higher than the costs of internal coordination, growth of the range of requests can give the supplier the economic advantage. Reducing both internal and external expenses of the transactions reduces the importance of the dimension that favors the production within the company's own limits. The impact of these two factors in practice manifests itself by reducing the size of the company, but also by reducing the average added value by company, which is easily experimentally verified in the countries that base their tax systems on added value. Correlation has been confirmed in many empirical analyses which deal with both global changes in these values during the past period, and the connection of these values with the information and communication technologies in the company.

Second main path of relation between size of the organization and application of information and communication technologies is automation. The automation is the concept driving economic development since the beginning of Industrial revolution, and most often defined as a decrease of participation of human labor in manufacture, and increase of machine automatic performance (Dulanović & Jaško 2006)³⁰ Although clearly separated from mechanization, automation has connections with it, and its roots are from mechanical automation which wider use started with Industrial revolution. Information and communication technologies have potentials of automation of different tasks that include handling and processing data. Since this description covers wide range of organizational tasks, it could be said that the majority of business can be improved by information and communication technologies intermediary, through the supporting tools, and great number of tasks would be totally or partly automated or eliminated in accordance with the reengineering approach (Hammer 1990)³¹.

Original appliance of computers in automation was based on text processing, calculations and reporting, as well as request processing. Dominate direction was automation of simple, routine tasks. That kind of automation routine is actually the most important part of any system based on computer – even of the systems which seems to deal with different matter. The systems based on the computer generally connecting specific routines on two levels: closed routines «hidden» in functions of inside applied program and opened routines which incorporate dialogues with users and structure their work. Potentials of the first group are often much smaller then the potentials of the other group, where attained synergy value of human factor extremely fast computer routine implementation.

Among numerous examples of automation aided by ICT is Fujicu factory which covering 20 000 square meters with 82 employees in daily shift, and only one operator in control room during night. His only task is to supervise industrial robots and automatic machine tools from control room. Traditional factory of the same size would employ ten times more workers. The automation of services gives even more possibilities. Particularly, works which as service include information – and this kind of works dominate in modern global economy, represent very good basis for automation. Administrative work in all companies, regardless of industry, is also automated. That complies with the claim that information and communication technologies significantly aid automation both in material

production (especially factory production), non-material production or services – every product or service which consists of information, data procession or data forwarding and internal trade administration (Groth, 1999)³² but also connected to outsourcing path automation of external and internal transactions as consequence of wider appliance of network models.

5. Conclusion

This research concludes that although theoretical and empirical researcher have shown that ICT reduces size of the company, medium and large companies adopt ICT with highest intensity. Size reduction is best described by the studies of Brynjolfsson et al. (1994).³³ The impact of information and communication technologies has also been studied through the total income of sales by companies that manufacture, which is expressed in the net sales price, including all the discounts. If the capital invested in informational technologies doubles, and all other values are observed as constants, the sale of the company reduces by 13%. Also there is significant correlation with 99,9% certainty, by which double value of the capital invested in informational technologies means 12% of reduction in added value of a company. This does not mean that investing in informational technologies negatively affects company's business. In the context of these results, we can recognize that this impact of the informational technology is directly linked to creating network models of organization and separating large vertically integrated companies into the smaller, more flexible entities.

It is logical that large enterprises which have potential to utilize ICT become leaders of ICT adoption, but also to decrease size toward medium enterprises, that keep high ICT adoption. Small organizations, at least in transitional countries, do not show high level of ICT use in average, but especially in smallest, micro organizations there can be extreme examples both of high and low ICT use, which is in such companies more dependent on other organizational factors.

References

1. Antonelli, D., Cassarino, I. and Villa, A. **Analysing collaborative demand and supply networks of SMEs**, International Journal of Networking and Virtual Organisations, 3/2, 2006, pp. 128 – 141
2. Arrow, K. J. **Information and Economic Behavior**, Boston, USA: Harvard University technical report, 1973
3. Audretsch, D. and Vivarelli, M. **Firm size and R&D spillovers: evidence from Italy**, Small Business Economics, 8, 1996, pp. 249–258
4. Blau, P. M. and Schoenherr, R. A. **The structure of organizations**, New York, USA: Basic books, 1971
5. Bloomfield, B. P. and Coombs, R. **Information technology, control and power: the centralization and decentralization debate revisited**, Journal of Management Studies, 29, 1992, pp. 459–484
6. Brynjolfsson, E., Malone, T., Gurbaxani, V. and Kambil, A. **Does Information Technology Lead to Smaller Firms?**, Management Science, 40(12), 1994, pp. 1628-1644

7. Cisler, S. **What's the Matter with ICTs**, chapter in Lovink, G. and Zehle, S. (eds.) "Incommunicado Reader: Information Technology for Everybody Else", Amsterdam, Netherlands: Institute of Network Cultures, 2005
8. Cudanov, M. **Projektovanje organizacije i IKT**, Beograd, Srbija: Zadužbina Andrejević, 2007
9. Cudanov, M., Jasko, O. and Jevtic, M. **Influence of Information and Communication Technologies on Decentralization of Organizational Structure**, Computer Science and Information Systems COMSIS, 6/1, 2009, pp. 93-108
10. Cudanov, M., Jasko, O. and Jevtic, M. **The Influence of Information and Communication Technologies on Organizational Performance**, InfoM - Journal of Information Technology and Multimedia Systems, Vol. 27, 2007, pp.1-9
11. Dulanovic, Z. and Jasko, O. **Osnovi organizacije poslovnih sistema**, Belgrade, Serbia: Faculty of organizational sciences, 2006
12. Galbraith, J. **Organizational Design**. Reading, MA, USA: Addison-Wesley, 1977
13. Groth, L. **Future organizational design – the scope for the IT based enterprise**, New York, USA: Wiley&Sons, 1999
14. Hammer, M. **Reengineering Work: Don't Automate, Obliterate**, Harvard Business Review, vol. 68 issue 4, 1990, pp.104-112
15. Kostic, K. **Softver za knjigovodstvo malog preduzeća**, Računovodstvo, 49/1-2, 2004, pp. 78-87
16. Kramer, W. J., Jenkins, B. and Katz, R. S. **The Role of the Information and Communications Technology Sector in Expanding Economic Opportunity**, Cambridge, Massachusetts: Kennedy School of Government, Harvard University, 2007
17. Lee, G. and Xia, W. **Organizational size and IT innovation adoption: A meta-analysis**, Information & Management, 43/8, 2006, pp. 975-985
18. Marmaridis, I. and Unhelkar, B. **Challenges in Mobile Transformations: A Requirements modelling perspective for Small and Medium Enterprises**, "Proceedings of the 4th International Conference on Mobile Business (ICMB 2005)", Sydney, Australia, 2005
19. Mintzberg, H. **Structure in 5's: A Synthesis of the Research on Organization Design**, Management Science, 26/3, 1980, pp. 322-341
20. Narula, R. **R&D Collaboration by SMEs: new opportunities and limitations in the face of globalization**, Technovation, 24/2, 2001, pp. 153-161
21. Osterman, P. **The Impact of Computers on the Employment of Clerks and Managers**, Industrial and Labor Relations Review, 39, 1986
22. Piore, M. and Sabel, C. **The Second Industrial Divide**, New York, USA: Basic Books, 1984
23. Trimi, S. **ICT for small and medium enterprises**, Service Business, 2/4, 2005, pp. 271-273
24. Van Belle, J. and Elis, J. **Open source software adoption by South African MSEs: barriers and enabler**, "Proceedings of the 2009 Annual Conference of the Southern African Computer Lecturers' Association", 2009, pp. 41-49
25. Van Dijk, B., Den Hertog, R., Menkveld, B. and Thurk, R. **Some new evidence on the determinants of large- and small-firm innovation**, Small Business Economics, 9, 1997, pp. 335-343
26. Weill, P. **The relationship between investment in information technology and firm performance: a study of the valve manufacturing sector**, Information Systems Research 3(4), 1992
27. Willcocks, P. and Lester, S. **Beyond the IT Productivity Paradox**, London, UK: John Wiley & Sons Ltd, 1999
28. Williamson, O. E. **Transaction-Cost Economics: The Governance of Contractual Relations**, The Journal of Law and Economics, Vol. 22, No. 2, 1979, pp. 233-261

29. * * * **Google Investor Relations**, available online http://investor.google.com/fin_data 2008.html, last accessed December 10th 2009
30. * * * **Official law publication**, Službeni list SRJ No.71/02 from 27.12. 2002

¹Mladen CUDANOV got his magister degree at Faculty of organizational sciences in 2006 and is working on his PhD thesis. Currently he is in assistant position at Faculty of Organizational Sciences, Organization of Business Systems department, University of Belgrade. He has been visiting for one semester as an assistant professor in joint programs of iVWA from Germany and Jiangsu College of Information Technology from Wuxi and Zhuhai City Polytechnics from Zhuhai in China. His major research interests are ICT and organizational design, restructuring of business systems and organizational change.

²Ondrej JASKO has graduated at the Faculty of Organizational Sciences, University of Belgrade, in 1989. He got his master degree at the same Faculty in 1995 and PhD degree at the same Faculty in 2000. Currently he is in professor position at Faculty of Organizational Sciences, Organization of Business Systems department, University of Belgrade. He was in position of vice dean, and in editing boards of our leading Journals in Management. His major research interests are organization theory and design, business systems, restructuring of business systems and organizational change.

³(Co)Author of the books: *Exploratory Domains of Econophysics. News EDEN I & II* (2009), *Statistica - Un mod stiintific de gandire* (2007), *Populatia lumii intre explozie si implozie demografica* (2006), *Cercetari si modelari de marketing. Metode cantitative in cercetarea pietei*(2005), *Universul preturilor si indicii interpret* (2001).

⁴ Van Dijk, B., Den Hertog, R., Menkveld, B. and Thurk, R. **Some new evidence on the determinants of large- and small-firm innovation**, *Small Business Economics*, 9, 1997, pp. 335-343

⁵Audretsch, D. and Vivarelli, M. **Firm size and R&D spillovers: evidence from Italy**, *Small Business Economics*, 8, 1996, pp. 249-258

⁶Narula, R. **R&D Collaboration by SMEs: new opportunities and limitations in the face of globalization**, *Technovation*, 24/2, 2001, pp 153-161

⁷Antonelli, D., Cassarino, I. and Villa, A. **Analysing collaborative demand and supply networks of SMEs**, *International Journal of Networking and Virtual Organisations*, 3/2, 2006, pp. 128 - 141

⁸Van Belle, J. and Elis, J. **Open source software adoption by South African MSEs: barriers and enablers**, in "Proceedings of the 2009 Annual Conference of the Southern African Computer Lecturers' Association", 2009, pp. 41-49

⁹ Cisler, S. **What's the Matter with ICTs**, in: Lovink, G. and Zehle, S. (eds.) "Incommunicado Reader: Information Technology for Everybody Else", Amsterdam, Netherlands: Institute of Network Cultures, 2005

¹⁰Marmaridis, I. and Unhelkar, B. **Challenges in Mobile Transformations: A Requirements modelling perspective for Small and Medium Enterprises**, "Proceedings of the 4th International Conference on Mobile Business (ICMB 2005)", Sydney, Australia, 2005

¹¹Trimi, S. **ICT for small and medium enterprises**, *Service Business*, 2/4, 2005, pp. 271-273

¹²Kramer, W.J., Jenkins, B. and Katz, R.S. **The Role of the Information and Communications Technology Sector in Expanding Economic Opportunity**, Cambridge, Massachusetts: Kennedy School of Government, Harvard University, 2007

¹³Blau, P.M. and Schoenherr, R.A. **The structure of organizations**, New York, USA: Basic books, 1971

¹⁴ Mintzberg H. **Structure in 5's: A Synthesis of the Research on Organization Design**, *Management Science*, 26/3, pp. 322-341, 1980

¹⁵Bloomfield, B. P. and Coombs, R. **Information technology, control and power: the centralization and decentralization debate revisited**, *Journal of Management Studies*, 29, 1992, pp. 459-484

¹⁶Cudanov, M. **Projektovanje organizacije i IKT**, Zadužbina Andrejević, Belgrade, Serbia, 2007

¹⁷* * * **Official law publication**, Službeni list SRJ No.71/02, 27.12.2002

¹⁸Cudanov, M., Jasko, O. and Jevtic, M. **Influence of Information and Communication Technologies on Decentralization of Organizational Structure**, *Computer Science and Information Systems COMSIS*, 6/1, 2009, pp. 93-108

-
- ¹⁹ Kostic, K. **Softver za knjigovodstvo malog preduzeća**, Računovodstvo, 49/1-2, 2004, pp. 78-87
- ²⁰ Lee, G. and Xia, W. **Organizational size and IT innovation adoption: A meta-analysis**, Information & Management, 43/8, 2006, pp. 975—985
- ²¹ Osterman, P. **The Impact of Computers on the Employment of Clerks and Managers**, Industrial and Labor Relations Review, 39, 1986
- ²² Arrow, K. J. **Information and Economic Behavior**, Boston, USA: Harvard University technical report, 1973
- ²³ Galbraith, J. **Organizational Design**, Reading, MA, USA: Addison-Wesley, 1977
- ²⁴ Piore, M. and Sabel, C. **The Second Industrial Divide**, New York, USA: Basic Books, 1984
- ²⁵ Weill, P. **The relationship between investment in information technology and firm performance: a study of the valve manufacturing sector**, Information Systems Research 3(4), 1992
- ²⁶ Willcocks, P. and Lester, S. **Beyond the IT Productivity Paradox**, London, UK: John Wiley & Sons Ltd, 1999
- ²⁷ Cudanov, M., Jasko, O. and Jevtic, M. **The Influence of Information and Communication Technologies on Organizational Performance**, InfoM - Journal of Information Technology and Multimedia Systems, Vol. 27, 2007, pp.1-9
- ²⁸ * * * **Google Investor Relations**, available online http://investor.google.com/fin_data2008.html last accessed December 10th 2009
- ²⁹ Williamson, O.E. **Transaction-Cost Economics: The Governance of Contractual Relations**, The Journal of Law and Economics, Vol. 22, No. 2, 1979, pp. 233-261
- ³⁰ Dulanovic, Z. and Jasko, O. **Osnovi organizacije poslovnih sistema**, Belgrade, Serbia: Faculty of organizational sciences, 2006
- ³¹ Hammer, M. **Reengineering Work: Don't Automate, Obliterate**, Harvard Business Review, vol. 68 issue 4, 1990, pp.104-112
- ³² Groth, L. **Future organizational design – the scope for the IT based enterprise**, New York, USA: Wiley&Sons, 1999
- ³³ Brynjolfsson E., Malone, T., Gurbaxani, V. and Kambil, A. **Does Information Technology Lead to Smaller Firms?**, Management Science, 40(12), 1994, pp. 1628-1644

MODELING THE RELIABILITY OF INFORMATION MANAGEMENT SYSTEMS BASED ON MISSION SPECIFIC TOOLS SET SOFTWARE

Cezar VASILESCU

Associate Professor, Regional Department of Defense Resources Management Studies,
National Defense University, Bucharest, Romania

E-mail: caesarv@crmra.ro



Abstract: *The operational environments in which information management systems operate determine the existence of complex situations. Consequently, the command and control flow can take different paths, which involve different “sets” of activities. Each of those activities is associated with a specific software application set, known as Application Software Tools (ASTs). An operational profile represents a sequence of specific processing of distinct activities (from a functional point of view), based on specific Application Software Tools and with a certain time limit interval. Each operational profile has associated a probability of occurrence.*

Each activity is performed during a specified period of time, with specific sets of ASTs. Totality resulting AST specification due to operational profiles crowd formed a mission specific software application system, also known as a Mission Specific Tools Set (MSTS). Each MSTS’s element fulfill functions that meet the corresponding command and control activities, found in the form of lists of features of the system operational profile.

The aim of this paper is to present an original MSTS reliability model, which combines the modelling approach based on operational profiles with Rome Research Laboratory software reliability modeling methodology. In this way, it was realized a dual representation of application set’s reliability that quantifies its level of reliability and also the associated weights of each application. The final goal was to offer an adequate basis for the process of reliability growth.

This paper is also going to provide a calculus example of MSTS system reliability using a representative U.S. Navy C4ISR system’s combat action (Time Critical Targeting). The case study demonstrates the validity and the usefulness of the model in order to increase the system’s reliability.

Key words: *Reliability modeling; Increase of software applications reliability; Operational profiles; Application software tools; Mission Specific Tools Set*

Introduction

Information management systems realize the processing of specific information necessary to conduct modern battlefield complex command and control activities, in order to ensure the success in battle. For mission-oriented software development it is necessary the modularization of the command and control activities and sub activities.

Generally, the operational profile can be defined as a quantitative characterization of the software usage, depending on the input space values. A profile consists of an independent possibilities set and their associated occurrence probabilities [6]¹.

The operational environments in which information management systems operate determine the existence of complex situations, characterized by a great diversity of information, inputs, actualization operations etc. Consequently, the command and control flow can take different paths, which involve different "sets" of activities. Each of those activities is associated with a specific software application set, known as *Application Software Tools (ASTs)*.

Speaking about information management systems, an operational profile represents a sequence of specific processing of distinct activities (from a functional point of view), based on specific Application Software Tools and with a certain time limit interval. Each operational profile has associated a probability of occurrence.

Each activity is performed during a specified period of time, with specific sets of ASTs. Totality resulting AST specification due to operational profiles crowd formed a mission specific software application system, also known as a *Mission Specific Tools Set (MSTS)*.

Each MSTS's element fulfill functions that meet the corresponding command and control activities, found in the form of lists of features of the system operational profile.

Calculation of MSTS system reliability will be subject to of following paragraph.

Calculation of MSTS system reliability

MSTS system reliability prediction and growth requires a dual core computing. This approach is driven by the possibility of joint activities under different distinct operational profiles.

The calculation relations are:

$$R_{MSTS} = \sum_{k=1}^{N_P} p_k R_k \quad (1)$$

in which

p_k - Occurrence probability of the k operational profile;

R_k - Reliability of the k operational profile;

N_P - Number of operational profiles.

The first relationship is based on the fact that each operational profile is associated with an occurrence probability [2].

Notation

$\alpha = \{\alpha_i; i = 1, N_\alpha\}$ = the set of MST activities;

$\alpha(k) = \{\alpha_i \in \alpha; \alpha_i \text{ belongs to the } k \text{ profile}\}$, ranked in ascending order of execution in the profile;

$\varphi = \{AST : AST \text{ is an instrument of the MSTs}\}$;

$\varphi(\alpha_i) = \{AST \in \varphi : AST \text{ serves } \alpha_i\}$;

AST = specific software application sets.

Then

$$R_k = \prod_{\alpha_i \in \alpha(k)} R_{\alpha_i}(t''_{\alpha_i} - t'_{\alpha_i}) \quad (2)$$

in which

t'_{α_i} = the beginning moment of activity α_i ;

t''_{α_i} = the ending moment of activity α_i .

and where

$$R_{\alpha_i}(t''_{\alpha_i} - t'_{\alpha_i}) = \prod_{AST \in \varphi(\alpha_i)} R_{AST}(t''_{\alpha_i} - t'_{\alpha_i}) \quad (3)$$

where

$$R_{AST}(t''_{\alpha_i} - t'_{\alpha_i}) = e^{-\lambda_{AST}(t''_{\alpha_i} - t'_{\alpha_i})} \quad (4)$$

$$\lambda_{AST}(t''_{\alpha_i} - t'_{\alpha_i}) = \sum \lambda_{AST} \quad (4bis)$$

The second calculation relation of MSTs system reliability is:

$$R_{MSTS}^{dual} = \prod_{k=1}^{N_p} R_k \quad (1 \text{ dual})$$

can be transformed as

$$R_{MSTS}^{dual} = \prod_{AST \in \varphi} R_{AST}^* \quad (5)$$

in which R_{AST}^* is the product of all factors in the formula (1 dual) that correspond to the same AST.

Note:

$R_{AST}^* = R_{AST}$ in case the AST appears one time in the formula (1 dual) and

$R_{AST}^* \neq R_{AST}, R_{AST}^* < R_{AST}$ otherwise.

This dualism is needed when profiles include joint activities. Using the first formula for calculating the reliability R_{MSTS} (in which components may occur several times) can provide the specification requirements for MSTs system reliability assessment and correspondence with the reliability requirements [1].

Also, MSTS system reliability calculation using the second relation (R_{MSTS}^{dual}) provides the possibility to organize the process of reliability growth, to meet the requirements specified. Thus in the calculation of reliability can be calculated weights (Π_{AST}) associated with AST and determined their influence.

$$\Pi_{AST} = \ln R_{AST}^* / \ln R_{MSTS}^{dual} \quad (6)$$

followed by the increasing ordering of the resulting string of weights $\{\Pi_{AST} : AST \in \varphi\}$, to highlight the order of priorities in addressing the growth of MSTS reliability. In this way can be highlighted MSTS components unsatisfactory in terms of reliability, so giving a good support to system software designers to eventually redesign it (if required) in the process of reliability growth.

In what follows, we present an example of calculating the MSTS system reliability, for the most common case in practical operation of the information management systems, in which under different operational profiles are common joint activities.

Case study

Depending on the nature, size and membership of the information management systems to a category of forces or other, command and control activities can have a high degree of specificity. In [3] there have been listed a number of typical command and control activities, and the general categories from which they belong. Also, in case of large information management systems analysis (e.g. national level), identification and analysis of all activities can be difficult.

For this reason, we calculate the MSTS system reliability [4] using for example one of the U.S. Navy C4ISR system's combat action. For this, it is necessary a brief overview of the C4ISR system and command and control activities related to combat action "Time Critical Targeting" [5].

U.S. Navy uses various systems against naval and air targets, with different C4ISR systems providing guidance. The flow of activities involved was analyzed, in order to optimize command and control activities, eliminate the overlapping functionality and ensure interoperability of systems.

Table 1. The main command and control sub activities, related with the combat action "Time Critical Targeting"

Current issue	Name
1.1.	Analysis of surveillance and reconnaissance data list
2.1.	Reconcile target combat priorities
2.2.	Determine sensor availability
2.3.	Task sensor
2.4.	Collect data
3.1.	Detect target
3.2.	Determine environment
3.3.	Tracking and positioning the target

Current issue	Name
3.4.	Identifying target
4.1.	Update target list
4.2.	Assess engagement capability
4.3.	Assign weapon-target Platform selection
4.4.	Update mission plans
4.5.	Perform TCT (time critical target) deconfliction
5.1.	Execute force order
5.2.	Support weapon flyout
5.3.	Fighting target
6.1.	Collect information on damage
6.2.	Damage information assessment
6.3.	Remove objective from target list

The flow of command and control activities related to combat action "Time Critical Targeting" (according to Table 1) is shown in Figure 1.

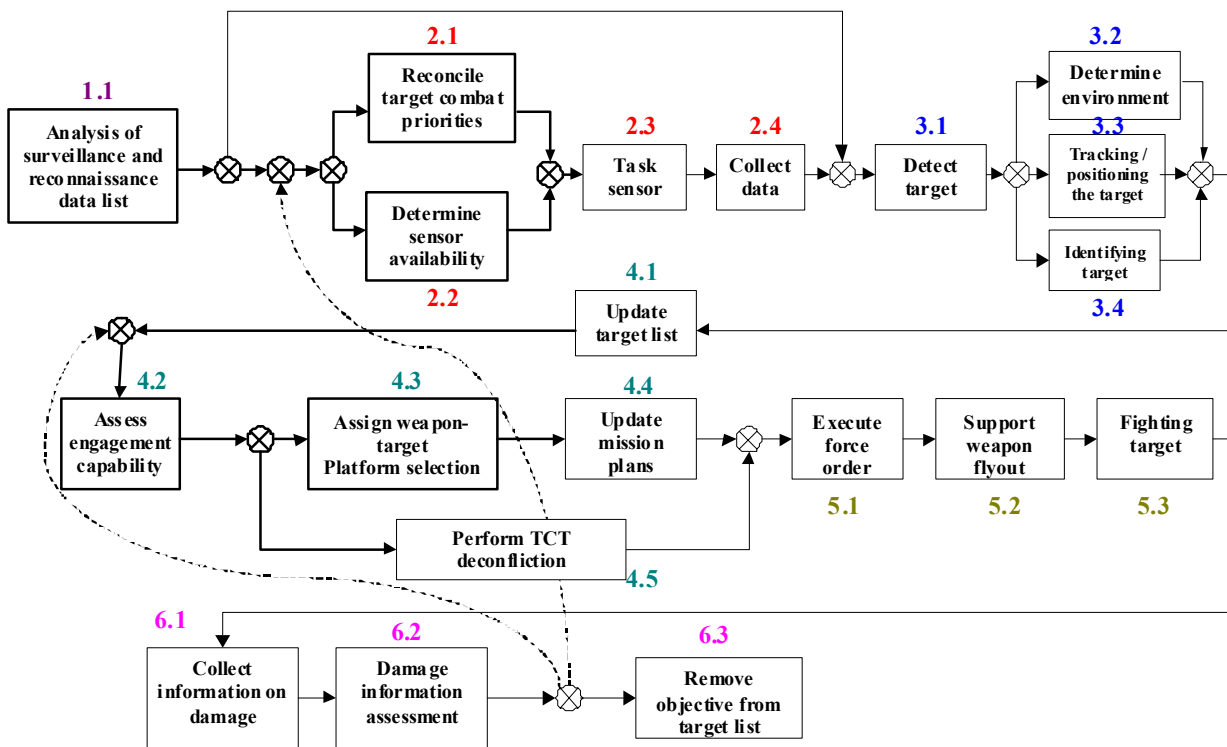


Figure 1. Scenario of C2 sub activities related to "Time Critical Targeting" combat action

AST names associated with sub activities are not relevant to the proposed goals. We present in terms of quantity the correlation between sub-C4ISR activities contained in Figure 1 and the number of software modules providing support to their deployment (Figure 2).

Typically, each operational profile of C4ISR activities is a chain of sequential actions. Application software tools sets are executed sequentially and/or competitive (exits a set representing the input for another set).

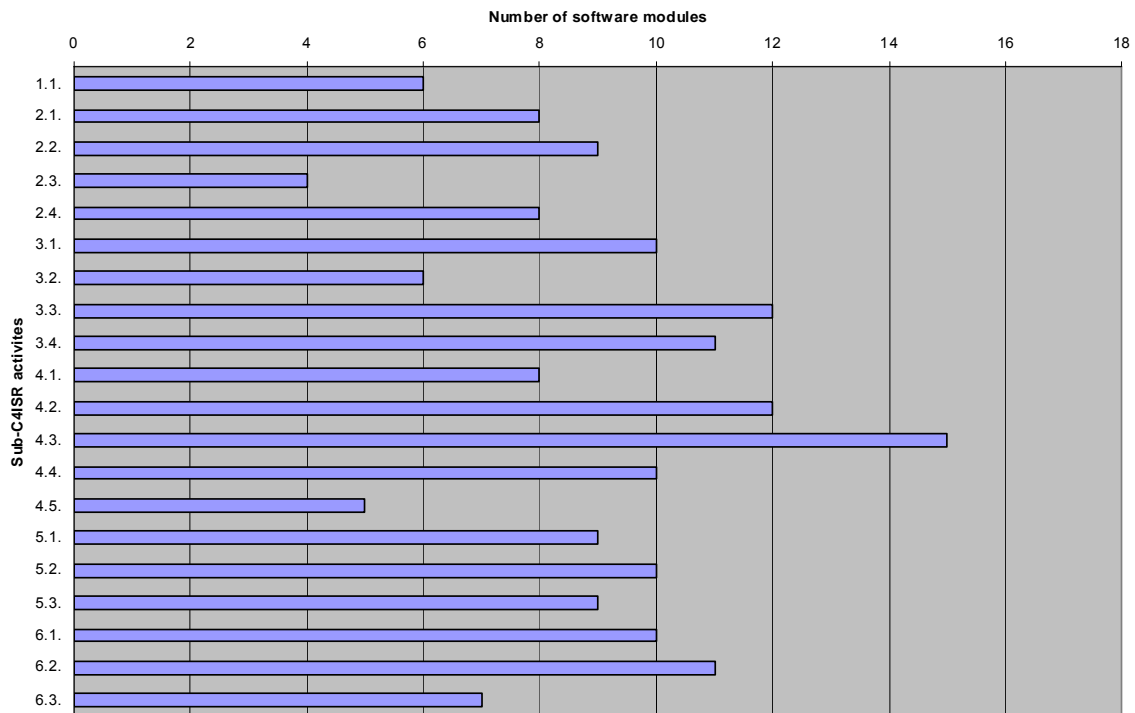


Figure 2. Graph of the number of software modules providing support to consisting sub-activities of “Time Critical Targeting” activity

We analyze the scenario of C4ISR sub activities related to “Time Critical Targeting” combat action (figure 1) to determine the operational profiles [4]. As a working hypothesis, we consider the entry of only one aircraft in the system (potential target) and use those numbers to each activity according to figure. The data related with operation of system’s software modules (values estimated for failure rates by type of software modules and times of activation, ie completion) were altered to serve for illustration purposes.

Step 1

Determine operational profiles (sequences of activities):

- *profile 1* (target entry into the system, fight and destroy it)
 (1.1) → (2.1) → (2.2) → (2.3) → (2.4) → (3.1) → (3.2) → (3.3) → (3.4) → (4.1) → (4.2) → (4.3) → (4.4) → (4.5) → (5.1) → (5.2) → (5.3) → (6.1) → (6.2) → (6.3)
- *profile 2* (target already challenged but still undamaged)
 (4.2) → (4.3) → (4.4) → (4.5) → (5.1) → (5.2) → (5.3) → (6.1) → (6.2) → (6.3)
- *profile 3* (target already challenged, still undamaged and emerged from the initial radar surveillance sector)
 (2.1) → (2.2) → (2.3) → (2.4) → (3.1) → (3.2) → (3.3) → (3.4) → (4.1) → (4.2) → (4.3) → (4.4) → (4.5) → (5.1) → (5.2) → (5.3) → (6.1) → (6.2) → (6.3)

Each C4ISR activity is done through a variable number of specific sets of software applications (AST). In turn, each AST consists of a variable number of independent software modules executed competitively (Table 2), whose characteristics are presented in Table 3.

Table 2. Correspondence between C4ISR activities, specific sets of software applications and number of software modules

C4ISR activities	Specific sets of software applications (AST)	Number of software modules
1	1.1.	6
2	2.1.	8
	2.2.	9
	2.3.	4
	2.4.	8
3	3.1.	10
	3.2.	6
	3.3.	12
	3.4.	11
4	4.1.	8
	4.2.	12
	4.3.	15
	4.4.	10
	4.5.	5
5	5.1.	9
	5.2.	10
	5.3.	9
6	6.1.	10
	6.2.	11
	6.3.	7

Step 2

We calculate for each AST the average failure rate and the reliability during operation.

We present detailed calculations for AST 1.1 and AST 2.1, following that for others to mention only the final results.

The average failure rate for AST is calculated using the equation:

$$\lambda_{AST} = \sum_{i=1}^m \lambda_{AST_i} ,$$

where

m = number of competitive active software modules corresponding to AST

$$\lambda_{AST1.1} = \sum_{i=1}^m \lambda_{AST_i} = (3 + 6 + 2 + 8 + 4 + 8) \times 10^{-5} = 0,00031 \text{ hours}^{-1}$$

$$\lambda_{AST2.1} = \sum_{i=1}^m \lambda_{AST_i} = (3 + 6 + 3 + 7 + 8 + 7 + 8 + 3) \times 10^{-5} = 0,00045 \text{ hours}^{-1}$$

Reliability function will be:

$$R_{AST1.1} = e^{-\lambda_{AST1.1}} = e^{-0,00031} = 0,99969$$

$$R_{AST2.1} = e^{-\lambda_{AST2.1}} = e^{-0,00045} = 0,99955$$

Table 4 presents values of average failure rates and reliability of specific application software sets.

Table 3. Characteristics of software modules sequentially active

AST	/ Types of software modules	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1.1	Activation time	0	45	200	300	800	900	-	-	-	-	-	-	-	-	-
	Completion time	45	200	300	800	900	1000	-	-	-	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	3	6	2	8	4	8	-	-	-	-	-	-	-	-	-
2.1	Activation time	0	50	100	250	400	600	750	980	-	-	-	-	-	-	-
	Completion time	50	100	250	400	600	750	980	1200	-	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	3	6	3	7	8	7	8	3	-	-	-	-	-	-	-
2.2	Activation time	0	55	100	150	300	500	650	800	1050	-	-	-	-	-	-
	Completion time	55	100	150	300	500	650	800	1050	1150	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	5	1	2	5	7	4	8	3	7	-	-	-	-	-	-
2.3	Activation time	0	45	200	300	-	-	-	-	-	-	-	-	-	-	-
	Completion time	45	200	300	600	-	-	-	-	-	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	2	8	4	5	-	-	-	-	-	-	-	-	-	-	-
2.4	Activation time	0	70	160	250	450	600	900	1000	-	-	-	-	-	-	-
	Completion time	70	160	250	450	600	900	1000	1200	-	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	3	6	7	2	1	8	7	4	-	-	-	-	-	-	-
3.1	Activation time	0	115	200	300	500	700	1000	1150	1240	1350	-	-	-	-	-
	Completion time	115	200	300	500	700	1000	1150	1240	1350	1500	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	1	2	8	5	5	3	6	8	8	9	-	-	-	-	-
3.2	Activation time	0	85	200	400	700	900	-	-	-	-	-	-	-	-	-
	Completion time	85	200	400	700	900	1000	-	-	-	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	2	4	4	2	5	7	-	-	-	-	-	-	-	-	-
3.3	Activation time	0	50	150	300	450	700	860	940	1025	1200	1350	1420	-	-	-
	Completion time	50	150	300	450	700	860	940	1025	1200	1350	1420	1550	-	-	-
	Failure rate (x10 ⁻⁵)	1	8	2	5	6	6	8	8	8	2	3	4	-	-	-
3.4	Activation time	0	85	150	250	500	600	850	930	1020	1250	1450	-	-	-	-
	Completion time	85	150	250	500	600	850	930	1020	1250	1450	1590	-	-	-	-
	Failure rate (x10 ⁻⁵)	2	2	3	7	5	6	1	8	5	8	7	-	-	-	-
4.1	Activation time	0	50	100	200	400	650	800	900	-	-	-	-	-	-	-
	Completion time	50	100	200	400	650	800	900	1050	-	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	3	5	6	5	3	4	2	7	-	-	-	-	-	-	-
4.2	Activation time	0	100	150	300	450	700	840	930	1000	1150	1320	1450	-	-	-
	Completion time	100	150	300	450	700	840	930	1000	1150	1320	1450	1600	-	-	-
	Failure rate (x10 ⁻⁵)	3	3	3	4	8	6	2	4	3	7	4	1	-	-	-
4.3	Activation time	0	75	130	250	400	500	740	820	1000	1090	1230	1310	1500	1600	1690
	Completion time	75	130	250	400	500	740	820	1000	1090	1230	1310	1500	1600	1690	1820
	Failure rate (x10 ⁻⁵)	2	4	3	5	8	2	3	7	6	6	6	4	3	2	9
4.4	Activation time	0	95	200	350	600	900	1100	1260	1450	1600	-	-	-	-	-
	Completion time	95	200	350	600	900	1100	1260	1450	1600	1800	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	2	1	2	3	6	7	8	6	4	4	-	-	-	-	-
4.5	Activation time	0	65	200	500	800	-	-	-	-	-	-	-	-	-	-
	Completion time	65	200	500	800	900	-	-	-	-	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	3	3	5	6	2	-	-	-	-	-	-	-	-	-	-
5.1	Activation time	0	100	200	300	400	500	750	850	1000	-	-	-	-	-	-
	Completion time	100	200	300	400	500	750	850	1000	1200	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	8	3	5	2	7	6	5	4	4	-	-	-	-	-	-
5.2	Activation time	0	115	200	350	800	900	1100	1300	1780	2000	-	-	-	-	-
	Completion time	115	200	350	800	900	1100	1300	1780	2000	2300	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	1	3	3	6	8	9	2	6	6	4	-	-	-	-	-
5.3	Activation time	0	100	300	500	700	900	1000	1200	1290	-	-	-	-	-	-
	Completion time	100	300	500	700	900	1000	1200	1290	1500	-	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	6	2	3	8	3	5	5	5	6	-	-	-	-	-	-
6.1	Activation time	0	55	120	450	670	800	895	975	1056	1170	-	-	-	-	-
	Completion time	55	120	450	670	800	895	975	1056	1170	1300	-	-	-	-	-
	Failure rate (x10 ⁻⁵)	2	3	6	3	1	2	6	7	8	3	-	-	-	-	-
6.2	Activation time	0	105	200	350	600	800	980	1100	1200	1350	1440	-	-	-	-
	Completion time	105	200	350	600	800	980	1100	1200	1350	1440	1565	-	-	-	-
	Failure rate (x10 ⁻⁵)	3	3	4	3	4	3	2	6	6	8	5	-	-	-	-

AST	/ Types of software modules	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6.3	Activation time	0	100	200	330	500	700	860	-	-	-	-	-	-	-	-
	Completion time	100	200	330	500	700	860	1200	-	-	-	-	-	-	-	-
	Failure rate ($\times 10^{-5}$)	5	6	3	4	1	7	8	-	-	-	-	-	-	-	-

Table 4. Values of average failure rates and reliability of specific application software sets

AST	λ_{AST}	R_{AST}
1.1	0,00031	0,99969
2.1	0,00045	0,99955
2.2	0,00042	0,99958
2.3	0,00019	0,99981
2.4	0,00038	0,99962
3.1	0,00055	0,99945
3.2	0,00024	0,99976
3.3	0,00061	0,99939
3.4	0,00054	0,99946
4.1	0,00035	0,99965
4.2	0,00048	0,99952
4.3	0,00070	0,99930
4.4	0,00043	0,99957
4.5	0,00019	0,99981
5.1	0,00044	0,99956
5.2	0,00048	0,99952
5.3	0,00043	0,99957
6.1	0,00041	0,99959
6.2	0,00047	0,99953
6.3	0,00034	0,99966

Step 3

We calculate the reliability of C4ISR activities R_{α_i} .

$$R_{\alpha_i} = \prod_{AST \in \varphi(\alpha_i)} R_{AST}$$

$$R_{\alpha_1} = R_{AST1.1} = 0,99969$$

$$R_{\alpha_2} = R_{AST2.1} \times R_{AST2.2} \times R_{AST2.3} \times R_{AST2.4} = 0,99865$$

$$R_{\alpha_3} = R_{AST3.1} \times R_{AST3.2} \times R_{AST3.3} \times R_{AST3.4} = 0,99806$$

$$R_{\alpha_4} = R_{AST4.1} \times R_{AST4.2} \times R_{AST4.3} \times R_{AST4.4} \times R_{AST4.5} = 0,99785$$

$$R_{\alpha_5} = R_{AST5.1} \times R_{AST5.2} \times R_{AST5.3} = 0,99865$$

$$R_{\alpha_6} = R_{AST6.1} \times R_{AST6.2} \times R_{AST6.3} = 0,99878$$

Step 4

The reliability of operational profiles R_{pk} is:

$$R_1 = \prod R_{\alpha_i} = R_{\alpha_1} \times R_{\alpha_2} \times R_{\alpha_3} \times R_{\alpha_4} \times R_{\alpha_5} \times R_{\alpha_6} = 0,991625$$

$$R_2 = R_{\alpha_4} \times R_{\alpha_5} \times R_{\alpha_6} = 0,995291$$

$$R_3 = R_{\alpha_2} \times R_{\alpha_3} \times R_{\alpha_4} \times R_{\alpha_5} \times R_{\alpha_6} = 0,991933$$

Consider the following values for the operational profiles' probability of occurrence p_k :

$$p_1 = 0,75$$

$$p_2 = 0,15$$

$$p_3 = 0,10$$

Step 5

MSTS reliability is:

$$R_{MSTS} = \sum_{k=1}^{N_p} p_k R_k = \sum_{k=1}^3 p_k R_k = p_1 R_1 + p_2 R_2 + p_3 R_3$$

$$R_{MSTS} = (0,75 \times 0,991625) + (0,15 \times 0,995291) + (0,10 \times 0,991933) = 0,992206$$

If using formula (1 dual) for MSTS system's reliability calculation, we can rewrite step 5, as follows:

Step 5 (dual)

MSTS reliability is:

$$R_{MSTS}^{dual} = \prod_{k=1}^{N_p} R_k = \prod_{k=1}^3 R_k = R_1 \times R_2 \times R_3 = 0,978994$$

Also, there is a new step:

Step 6

We calculate the weights \prod_{AST} associated with each AST using the formula:

$$\prod_{AST} = \ln R_{AST}^* / \ln R_{MSTS}^{dual}$$

We present detailed calculations for AST 1.1 and AST 2.1 associated weights, following that for others to mention only the final results.

Table 5 present values of weights associated to each specific application software set.

$$\prod_{AST1.1} = \ln R_{AST1.1} / \ln R_{MSTS}^{dual} = \ln(0,99969) / \ln(0,978994) = 0,036510$$

$$\prod_{AST2.1} = \ln R_{AST2.1} / \ln R_{MSTS}^{dual} = \ln(0,99955) / \ln(0,978994) = 0,053003$$

$$\prod_{AST2.2} = \ln R_{AST2.2} / \ln R_{MSTS}^{dual} = \ln(0,99958) / \ln(0,978994) = 0,049470$$

Table 5. The values of weights associated with application software sets.

AST	R_{AST}	\prod_{AST}
1.1	0,99969	0,036510
2.1	0,99955	0,053003
2.2	0,99958	0,049470

AST	R_{AST}	\prod_{AST}
2.3	0,99981	0,022378
2.4	0,99962	0,044758
3.1	0,99945	0,064785
3.2	0,99976	0,028265
3.3	0,99939	0,071855
3.4	0,99946	0,063608
4.1	0,99965	0,041223
4.2	0,99952	0,056538
4.3	0,99930	0,082460
4.4	0,99957	0,050648
4.5	0,99981	0,022378
5.1	0,99956	0,051825
5.2	0,99952	0,056538
5.3	0,99957	0,050648
6.1	0,99959	0,048290
6.2	0,99953	0,055360
6.3	0,99966	0,040045

We execute the decreasing ordering of the weights result string.

$$(\prod_{AST})_{AST \in \varphi} = (\prod_{4,3}, \prod_{3,3}, \prod_{3,1}, \prod_{3,4}, \prod_{4,2}, \prod_{5,2}, \prod_{6,2}, \prod_{2,1}, \prod_{5,1}, \prod_{4,4}, \prod_{5,3}, \prod_{2,2}, \prod_{6,1}, \prod_{2,4}, \prod_{4,1}, \prod_{6,3}, \prod_{1,1}, \prod_{3,2}, \prod_{2,3}, \prod_{4,5})$$

The conclusion offered by the decreasing ordering of this string is that AST4.3 and AST3.3 have the largest weight (influence) on the reliability of the whole, any redesign of the software modules that compose AST4.3 and AST3.3 being highly recommended in the reliability increasing process.

References

1. Ghita, A. and Ionescu, V. **Metode de calcul în fiabilitate**, Military Technical Academy Publishing House, Bucharest, 1996
2. Musa, J.D. **Operational Profiles in Software Reliability Engineering**, IEEE Software Magazine, March, 1993
3. Vasilescu, C. and Ghita, A. **Contributions to the field of C4ISR Systems Reliability Modeling**, „Proceedings of the XXXIst Annual Scientific Communications Session with international attendance”, Military Technical Academy, Bucharest, 3-4 November 2005
4. Vasilescu, C. **Prezentare generală a sistemelor de comandă și control**, National Scientific Communications Session, Air Force Academy, Brasov, November 2000
5. Vasilescu, C. **Probleme actuale în fiabilitatea sistemelor de comanda și control din Fortele Aeriene**, Editura Universitatii Transilvania din Brasov, 2009, pp. 236-245
6. * * * **DoD Architecture Framework version 1.0 Deskbook**, DoD Architecture Working Group, August 2003
7. * * * **System and Software Reliability Assurance Notebook**, produced for Rome Laboratory, New York, 1997

¹ Codification of references:

[1]	Ghita, A. and Ionescu, V. Metode de calcul în fiabilitate , Military Technical Academy Publishing House, Bucharest, 1996
[2]	Musa, J.D. Operational Profiles in Software Reliability Engineering , IEEE Software Magazine, March, 1993
[3]	Vasilescu, C. Prezentare generală a sistemelor de comandă și control , National Scientific Communications Session, Air Force Academy, Brasov, November 2000
[4]	Vasilescu, C. and Ghita, A. Contributions to the field of C4ISR Systems Reliability Modeling , „Proceedings of the XXXIst Annual Scientific Communications Session with international attendance”, Military Technical Academy, Bucharest, 3-4 November 2005
[5]	* * * DoD Architecture Framework version 1.0 Deskbook , DoD Architecture Working Group, August 2003
[6]	* * * System and Software Reliability Assurance Notebook , produced for Rome Laboratory, New York, 1997
[7]	Vasilescu, C. Probleme actuale în fiabilitatea sistemelor de comanda si control din Fortele Aeriene , Editura Universitatii Transilvania din Brasov, 2009, pp. 236-245

DATA MINING INTO THE WEBSITES OF MANAGEMENT INSTITUTES USING BINARY REPRESENTATION

Hemanta SAIKIA

Department of Business Administration,
Assam University, Assam, India

E-mail: h.saikia456@gmail.com

Dibyoyoti BHATTACHARJEE

Department of Business Administration,
Assam University, Assam, India

E-mail: dibyoyoti.bhattacharjee@gmail.com



Abstract: A similarity index is developed in this paper to measure the resemblance of information contained in the websites of several management institutes of India. The data matrix pertaining to information contents of the different websites is populated using indicator variables. A Pair Similarity Index (PSI), for non-mutually exclusive cases, is proposed that can measure the similarity between websites through pairs of observations. A comparison of the proposed similarity index with one such existing index is also done.

Key words: Binary representation; data mining; website comparison; similarity index

1. Introduction

The World Wide Web has played an important role in presenting the data, even from geographically distant locations, easily accessible to users all over the world. A website is a collection of web pages, consisting of text and images that provide information about a particular topic or organization, twenty four hours a day and seven days a week (Bhattacharjee and Gupta, 2008). Today, it's a big challenge for management institutes to stay upgraded in global educational environment. Most of the management institutes provide information about students, courses, faculty, staff and facilities available and other details through their websites and accordingly market themselves. All these information are useful for the students, guardians, scholars as they get a bird's eye view about the institute. Having a website helps the administration of any institute to provide information about their services namely admission, results, rules, placement, etc. and accordingly diminish their

work load to a greater extent.

In India, there are many government and privately run management institutes. Every year, these management institutes are ranked by All India Council of Technical Education (AICTE) based on the institutes Intellectual Capital, Admission and Placements, Infrastructure, Industry Interface and Governance, etc. The proposed study is based on information contained into the websites of 21 top management institutes that were ranked by AICTE in the year 2008. The information contained in the websites of the management institutes were classified into some categories and under each category many attributes are considered. If a particular information, is provided in the institute's websites then it is coded as "1" and otherwise "0". In the study independence between the categories are assumed. The main aim of this paper is to develop a Paired Similarity Index (PSI) to study the similarity between any two websites of the management institutes.

2. Objectives of the study

The objectives of the proposed study are as follows –

1. To develop a Paired Similarity Index (PSI) (for non-mutually exclusive cases) by extending an earlier work due to Erlich, Gelbard and Spiegler (2002).
2. To study the similarity of websites of management institutes of India by using the proposed PSI.

3. Methodology

The different management institutes considered for the study are as follows IIM Ahmedabad (IIMA), IIM Bangalore (IIMB), IIM Calcutta (IIMC), ISB Hyderabad (ISBH), IIM Lucknow (IIML), XLRI Jamshedpur (XLRIJ), FMS Delhi (FMSD), IIM Indore (IIMI), IIM Kozhikode (IIMK), IIFT Delhi (IIFTD), SP Jain Mumbai (SPJM), MDI Gurgaon (MDIG), JBIMS Mumbai (JBIMSM), NMIMS Mumbai (NMIMSM), IMT Ghaziabad (IMTG), NITIE Mumbai (NITIE), SIBM Pune (SIBMP), XIMB Bhubaneswar (XIMBB), TISS Mumbai (TISSM), IIT Mumbai (IITM) and IIT Delhi (IITD). Following the website of IIM Ahmedabad, the best management institute of India, as per AICTE ranking, the information contained into the websites is classified into eight categories viz,

1. Admission procedure
2. Library facilities
3. Students
4. Other facilities (Hostel, sports, etc)
5. Faculty search
6. Research and development
7. Alumni association
8. Placement

Under each of these categories many attributes are considered, details of which is provided in Appendix-A. The availability of information about any attribute, in a given website is expressed by an indicator variable. The relevant data was collected from the websites of the management institutes in the month of August, 2009.

4. Review of literature

A review about some works related to data mining tool using binary data can be found in storage and retrieval considerations of binary data base by Spiegler and Maayan (1985), Fayyad, Haussler and Stolorz (1996) (data classification), data clustering is given by Jain, Murty and Flynn (1999), Gelbard and Spiegler (2000) (data clustering). Erlich et al. (2002) developed a model for similarity and clustering by means of binary representation for mutually exclusive cases.

5. PSI for binary data

Erlich, Gelbard and Spiegler (2002) proposed a data mining method by means of binary representation for determining pair similarity index between any two entities. Here we have a collection of websites of management institutes. The information content in the websites is subdivided into some broad categories. Under each category we consider some attributes. Then for each category under each website, we construct a binary vector that represents the presence (1) or absence (0) of its attributes. In this context the measure of similarity as proposed by Erlich et al. can be explained as follows–

Suppose that for each website 'i' ($i=1, 2, \dots, n$) we have 'm' categories. For each category j ($j=1, 2, \dots, m$) we have p_j attributes. The value p_j is called as the domain size of the j^{th} category. They define the binary representation vector of length, $p = \sum_{j=1}^m p_j$ (the

length of domain category vector), for each website 'i' ($i=1, 2, \dots, n$) in the following way –

$$x_{ijk} = 1, \text{ if the information about the } k^{\text{th}} \text{ attribute belonging to the } j^{\text{th}} \text{ category is available in the } i^{\text{th}} \text{ website.}$$

$$= 0, \text{ otherwise}$$

where $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$ and $k = 1, 2, \dots, p_j$

The mutual exclusivity property for each category over its domain was assumed. Using binary representation, Erlich et al. (2002) defined a pair similarity index (PSI) is as follows –

$$PSI = PSI(i_1, i_2) = \frac{sa(i_1, i_2)}{m} \tag{1}$$

$$\text{where } m = \sum_{j=1}^m \sum_{k=1}^{p_j} x_{ijk} \text{ and } sa(i_1, i_2) = \sum_{j=1}^m \sum_{k=1}^{p_j} x_{i_1 j k} = \sum_{j=1}^m \sum_{k=1}^{p_j} x_{i_2 j k}$$

Now for each category j , if a website can attain maximum possible of its p_j domain values (i.e. when the mutually exclusivity property doesn't satisfied for each category over its domain) then the range of pair similarity index (PSI) given by Erlich et al. (2002) is greater than one (i.e. $PSI > 1$). If the value of PSI is greater than one then it is difficult to determine the similarity measure between websites of any two management institutes. Therefore, we cannot designate absolute similarity between any two websites in case of binary representation using (1). So we develop a new pair similarity index, as the ratio between the number of similar attribute values of any two websites and the length of the domain attribute vector to overcome the above mentioned difficulties for non-mutually exclusive cases. Thus, we redefine the PSI for any two websites i_1 and i_2 is as follows –

$$PSI = \frac{sa(i_1, i_2)}{p} \tag{2}$$

where $p = \sum_{j=1}^m p_j$ and $sa(i_1, i_2) = \sum_{j=i}^m \sum_{k=1}^{p_j} x_{i_1 j_k} = \sum_{j=1}^m \sum_{k=1}^{p_j} x_{i_2 j_k}$

Now, the similarity index range is becomes $0 \leq PSI \leq 1$. Where $PSI = 1$ denotes absolute similarity and $PSI = 0$ denotes absolute diversity between any two websites of the management institutes.

Example: Let us take the binary representation vectors for the management institute $i_1=IIMA$ and $i_2=IIMB$ from Appendix-B. In order to calculate the Paired Similarity Index for any two management institutes first we calculate $sa(i_1, i_2)$.

$$sa(i_1, i_2) = \sum_{j=1}^8 \sum_{\substack{k=1 \\ x_{1jk}=x_{2jk}}}^{41} x_{1jk} = 24 \tag{3}$$

and therefore using (2)

$$PSI = \frac{sa(i_1, i_2)}{p} = \frac{24}{41} = 0.585$$

Since, the value of PSI lies between 0 and 1 so this value of 0.585 indicates very negligible similarity between the websites of IIMA and IIMB.

Similarly, the PSI values for all the pairs of management institutes formed for the 21 management institutes were calculated. The results of the corresponding pair similarity index matrix can be seen in Appendix-C.

6. PSI and other similarity indexes (for non-mutually exclusive cases)

A comparison of the proposed Paired Similarity Index (PSI) with other similarity indexes used in binary representation viz, Hamming Distance (HD) proposed by Illingworth, Glaser and Pyle (1983) and Paired Attribute Distance (PAD) proposed by Gelbard and Spiegler (2000) are as follows.

6.1 Comparing with HD: For two binary vector b_1 and b_2 , of length p , the HD between two vectors is defined as –

$$HD(b_1, b_2) = b_1 \oplus b_2$$

where \oplus denotes the logical operation XOR (Exclusive OR)

Gelbard and Spiegler (2000) give the normalized index based on HD by S_{HD} and it's defined as –

$$S_{HD}(b_1, b_2) = 1 - \frac{HD(b_1, b_2)}{p} = \frac{p - HD(b_1, b_2)}{p}$$

where $0 \leq S_{HD}(b_1, b_2) \leq 1$

and $HD(b_1, b_2)$ is the number of 1's in the vector b_1 and b_2 .

However, Erlich et al. (2002), already proved that the normalized similarity index S_{HD} given by Gelbard and Spiegler (2000) gives an incorrect measure to study the similarities of any two websites of management institutes in binary representation.

6.2 Comparing with PAD: The PAD similarity index as described in Gelberd and Spiegler (2000) for two binary vectors b_1 and b_2 is given by –

$$PAD = \frac{2 Nb_1 b_2}{Nb_1 + Nb_2}$$

where Nb_1 = the number of 1's in b_1

Nb_2 = the number of 1's in b_2

$Nb_1 b_2$ = the number of 1's common to both b_1 and b_2

In our binary representation we may have

$$Nb_1 = Nb_2 = \sum_{j=1}^m \sum_{k=1}^{p_j} x_{1jk} = p, \text{ for all 1's}$$

For instance from Appendix-B we consider the category “Admission” which has seven attributes. The corresponding binary representation of two institutes IIMA and IIMC for “Admission” are as follows –

$$b_1 = 1110010$$

$$b_2 = 1111110$$

then,
$$PAD = \frac{2 * 4}{4 + 6} = \frac{8}{10} = 0.8$$

Thus, the range of PAD is $0 \leq PAD \leq 1$. Therefore, in case of non-mutually exclusive cases the range of PAD and PSI is similar to measure the similarity or dissimilarity between any two management institutes by means of binary representation. The PAD for all the management institutes formed for the 21 management institutes were calculated and the results of the corresponding PAD matrix can be seen in Appendix-D.

7. Graphical display of PSI and PAD

Figure 1 provides the graphical representation of the values of similarity indices obtained under PSI and PAD for different pairs of institutes with IIMA common in all the pairs.

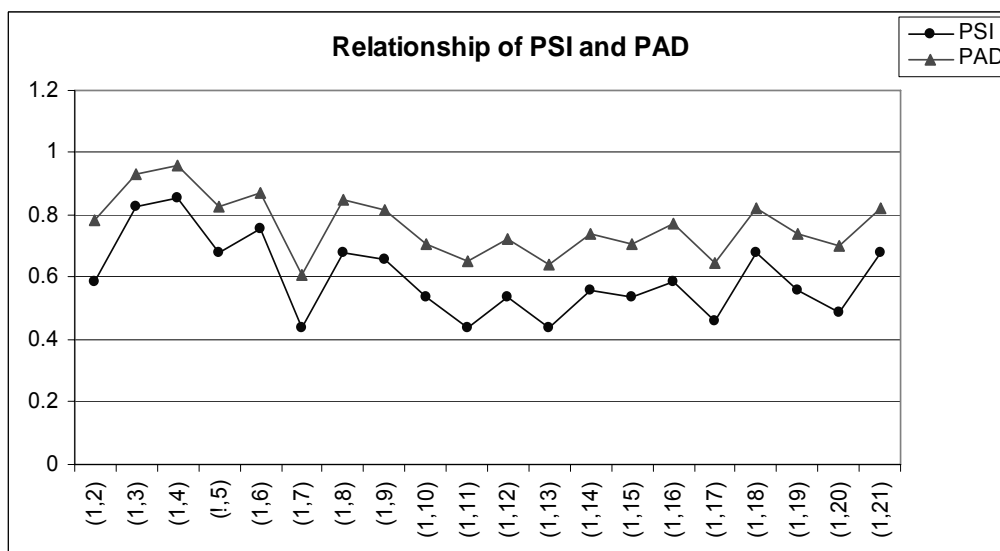


Figure 1. Line diagram showing difference between PSI and PAD

The graph shows that the results obtained from the PSI generally remains less than those obtained under PAD.

8. Results and findings

From the PSI matrix, it has been noticed that the value of PSI for the management institutes IIMA-ISBH and IIMC-ISBH are 0.853 and IIMA-IIMC is 0.829. Therefore, we can conclude that the information contained in the websites of the management institutes IIMA-ISBH and IIMC-ISBH are more similar among all other management institutes and IIMA-IMC inhabit second position in case of similarity measure. The maximum dissimilarity was noted between the websites of JBIMSM-IITM as their corresponding PSI value is 0.268. Also the study found that the information provided in the websites of the management institutes has no relation with the rank of the institutes as evident from Appendix-C and Appendix-D.

9. Conclusion

This study generates a new approach to measure similarity or dissimilarity by means of binary representation. However, the proposed paired similarity index, can handle a wide range of data types, continuous and multiple value domains. Handling of continuous data under this paired similarity index should be in categories. Deciding the number of categories is not a trivial problem by the choice of user. Also weights may be taken for different category and applied in this paired similarity index which will add relative importance of the categories to the proposed index. The index can find its application is several other disciplines of social science where similarity or dissimilarity needs to be measured.

References

1. Bhattacharjee, D. and Gupta, R. **A Meta Analytic Model for Comparing the Department of an Institute**, The Icfai University Journal of Computational Mathematics, Vol. 1(2), 2008, pp. 46-55
2. Erlich, Z., Gelbard, R. and Spiegler, I. **Data Mining by Means of Binary Representation: A Model for Similarity and Clustering**, Information Systems Frontiers, Vol. 4(2), 2002, pp. 187-197
3. Fayyad, U., Haussler, D. and Stolorz, P. **Mining scientific data**, Communications of the ACM, 1996, pp. 51 – 57
4. Gelbard, R. and Spiegler, I. **Hempel Raven Paradox: A Positive Approach to Cluster Analysis**, Computers and Operations Research, Vol. 27, 2000, pp. 305-320
5. Illingworth, V., Glaser, E. L. and Pyle, I. C. **Hamming distance**, Dictionary of Computing, Oxford University Press, 1983, pp.162-163
6. Jain, A.K., Murty, M.N. and Flynn, P.J. **Data Clustering: A Review**, ACM Computing Surveys, Vol. 31 (3), 1999, pp. 264 – 223
7. Spiegler, I. and Maayan, R. **Storage and retrieval considerations of binary data bases**, Information Processing & Management, Vol. 21(3), 1985, pp. 233-254

Appendix A

<p>1. Admission</p> <ul style="list-style-type: none"> • Prospectus • Doctoral Program • Full time Program • Part time Program • Fees per course • Contact Email • Information for foreign student 	<p>2. Library facilities</p> <ul style="list-style-type: none"> • Staff • Membership • Library layout • Rules & Regulations • Contact Email • Collection
<p>3. Students</p> <ul style="list-style-type: none"> • Role & Participation • Reservation • Financial Aid-program • Results • Fellowship • Student Union 	<p>4. Other facilities</p> <ul style="list-style-type: none"> • Hostel • Guest House • Medical • Sports • Award
<p>5. Faculty search</p> <ul style="list-style-type: none"> • Name & Designation • School & Department • Research & Publication • List of Teachers 	<p>6. Research & Development</p> <ul style="list-style-type: none"> • Faculty development program • Research & Publication • Management development program • Seminar/Workshop/Conference
<p>7. Alumni association</p> <ul style="list-style-type: none"> • Alumni relation • Activities • Alumni search criteria • Contact Email 	<p>8. Placement</p> <ul style="list-style-type: none"> • List of companies • Guidance • Brochure • Process • Contact Email

Appendix B

	ADMISSION						
	Prospectus	Doctoral program	Full time Program	Part time program	Fees per course	Contact E-mail	Information for foreign student
IIMA	1	1	1	0	0	1	0
IIMB	1	1	1	0	0	0	0
IIMC	1	1	1	1	1	1	0
ISBH	1	1	1	0	0	1	1
IIML	0	0	1	0	0	1	1
XLRIJ	1	1	1	1	0	0	1
FMSD	0	1	1	1	1	0	1
IIMI	1	1	1	0	0	1	1
IIMK	1	1	1	1	1	1	1
IIFTD	0	1	1	1	1	1	1

SPJM	0	1	1	0	1	1	0
MDIG	1	1	1	0	1	0	1
JBIMSM	0	1	1	1	1	1	0
NMIMSM	1	1	1	1	0	1	1
IMTG	1	1	1	1	1	1	0
NITIEM	1	1	1	0	1	1	0
SIBMP	0	0	1	0	1	1	1
XIMBB	1	1	1	1	1	1	1
TISSM	1	1	1	1	0	1	1
IITM	1	1	1	0	1	1	0
IITD	1	1	1	1	1	1	1

	LIBRARY					
	Staff	Membership	Library layout	Rules & Regulations	Contact E-mail	Collection
IIMA	1	1	1	1	1	1
IIMB	0	0	0	0	0	1
IIMC	1	1	1	1	1	1
ISBH	1	1	1	1	1	1
IIML	1	1	0	1	1	1
XLRIJ	1	1	1	1	1	1
FMSD	0	0	0	0	0	1
IIMI	0	1	0	0	1	1
IIMK	0	1	1	1	1	1
IIFTD	0	0	0	0	0	1
SPJM	0	0	0	0	0	1
MDIG	0	0	0	0	1	1
JBIMSM	0	0	0	0	0	1
NMIMSM	0	1	0	0	1	1
IMTG	0	1	0	1	1	1
NITIEM	0	0	0	0	1	1
SIBMP	0	0	0	0	0	0
XIMBB	1	1	0	1	1	1
TISSM	1	0	0	0	1	1
IITM	0	0	0	0	1	1
IITD	1	1	1	1	1	1

	PLACEMENT				
	List of companies	Guidance	Students profile or Brochure	Process	Contact E-mail
IIMA	1	1	1	1	1
IIMB	0	0	0	1	0
IIMC	0	1	1	1	1
ISBH	1	1	1	1	1
IIML	1	1	0	1	1
XLRIJ	1	0	1	1	1
FMSD	0	0	0	1	1
IIMI	1	1	1	0	1
IIMK	1	0	0	1	1
IIFTD	1	0	1	1	1
SPJM	1	0	1	1	1
MDIG	1	0	1	1	1
JBIMSM	1	0	0	1	1
NMIMSM	1	0	1	1	1
IMTG	1	0	0	0	1
NITIEM	1	0	1	1	1
SIBMP	1	0	1	1	1
XIMBB	0	0	1	1	1
TISSM	0	0	0	0	0

IITM	0	1	1	0	1
IITD	1	0	0	0	0

	RESEARCH AND DEVELOPMENT			
	Faculty development program	Research Publication	& Management development program	Seminar/workshop/conference
IIMA	1	1	1	1
IIMB	1	1	1	1
IIMC	1	1	1	1
ISBH	1	1	1	1
IIML	1	1	1	1
XLRIJ	1	1	1	1
FMSD	0	1	1	1
IIMI	1	1	1	1
IIMK	1	1	1	1
IIFTD	0	1	1	0
SPJM	0	1	0	1
MDIG	0	1	1	1
JBIMSM	0	0	0	1
NMIMSM	1	1	1	1
IMTG	0	1	1	1
NITIEM	0	1	1	1
SIBMP	0	1	0	1
XIMBB	0	1	1	1
TISSM	0	1	0	1
IITM	0	1	1	1
IITD	0	1	1	1

	STUDENTS					
	Role & participation	Reservation	Financial Aid program	Academic Result	Fellowship	Students union
IIMA	1	1	1	0	1	0
IIMB	1	0	1	0	1	1
IIMC	1	1	1	1	1	1
ISBH	1	0	1	0	1	1
IIML	1	0	0	0	0	1
XLRIJ	1	0	1	1	1	1
FMSD	1	1	0	1	0	1
IIMI	1	1	0	0	0	1
IIMK	1	1	1	0	1	0
IIFTD	1	1	0	0	0	1
SPJM	1	0	0	0	0	0
MDIG	1	0	0	1	0	0
JBIMSM	1	0	0	0	0	0
NMIMSM	1	0	0	0	0	1
IMTG	1	0	0	1	0	1
NITIEM	1	1	0	0	1	1
SIBMP	1	1	0	1	1	1
XIMBB	1	0	1	0	1	1
TISSM	1	1	1	0	1	1
IITM	1	0	0	0	1	0
IITD	1	1	0	0	1	1

	ALUMNI			
	Alumni relation	Activities	Alumni search criteria	Contact e-mail
IIMA	1	1	1	1



IIMB	1	1	1	1
IIMC	1	1	1	1
ISBH	1	1	1	1
IIML	1	1	1	1
XLRIJ	1	1	1	1
FMSD	1	1	1	1
IIMI	0	1	0	1
IIMK	0	0	0	0
IIFTD	1	1	1	1
SPJM	1	1	1	1
MDIG	0	1	1	1
JBIMSM	1	1	1	1
NMIMSM	0	0	0	0
IMTG	0	1	0	0
NITIEM	0	0	0	0
SIBMP	1	1	1	1
XIMBB	1	1	1	1
TISSM	1	1	0	1
IITM	0	0	0	0
IITD	1	1	1	1

	FACULTY			
	Name & Designation	School & Department	Research & publication	Teachers list
IIMA	1	1	1	1
IIMB	1	1	1	1
IIMC	1	1	1	1
ISBH	1	1	1	1
IIML	1	1	1	1
XLRIJ	1	1	1	1
FMSD	0	0	1	0
IIMI	1	1	1	1
IIMK	1	1	1	1
IIFTD	0	0	1	1
SPJM	0	0	1	1
MDIG	1	1	1	1
JBIMSM	1	0	0	1
NMIMSM	1	1	1	1
IMTG	1	1	1	1
NITIEM	1	1	1	1
SIBMP	0	0	0	1
XIMBB	1	1	1	1
TISSM	1	1	1	1
IITM	0	0	1	1
IITD	1	0	1	1

	FACILITIES				
	Hostel	Guest House	Medical	Sports	Award
IIMA	1	1	1	1	1
IIMB	1	1	0	1	1
IIMC	1	1	1	1	1
ISBH	1	1	1	1	1
IIML	1	1	0	1	0
XLRIJ	1	0	0	1	1
FMSD	1	0	0	1	1
IIMI	1	1	1	1	1
IIMK	1	1	0	1	0
IIFTD	1	1	0	1	1
SPJM	1	0	0	0	0

MDIG	1	0	0	1	0
JBIMSM	1	0	0	1	1
NMIMSM	1	0	1	0	1
IMTG	1	0	0	1	1
NITIEM	1	1	0	1	1
SIBMP	1	0	0	1	1
XIMBB	1	0	0	1	0
TISSM	1	0	1	1	0
IITM	1	1	1	1	0
IITD	1	1	1	1	0

Appendix C. PSI Matrix

	IIMA	IIMB	IIMC	ISBH	IIML	XLRIJ	FMSD	IIMI	IIMK	IIFTD	SPJM	MDIG	JBIMSM	NMIMSM	IMTG	NITIEM	SIBMP	XIMBB	TISSM	IITM	
IIMB	0.585	-																			
IIMC	0.829	0.609	-																		
ISBH	0.853	0.609	0.853	-																	
IIML	0.682	0.487	0.658	0.707	-																
XLRIJ	0.756	0.536	0.804	0.804	0.634	-															
FMSD	0.439	0.414	0.536	0.463	0.414	0.512	-														
IIMI	0.682	0.487	0.658	0.707	0.561	0.609	0.414	-													
IIMK	0.658	0.463	0.658	0.658	0.536	0.634	0.39	0.56	-												
IIFTD	0.536	0.439	0.585	0.56	0.487	0.536	0.512	0.512	0.463	-											
SPJM	0.439	0.341	0.439	0.439	0.39	0.463	0.365	0.365	0.341	0.439	-										
MDIG	0.536	0.439	0.56	0.56	0.487	0.585	0.439	0.512	0.487	0.463	0.414	-									
JBIMSM	0.439	0.365	0.463	0.439	0.39	0.439	0.39	0.365	0.365	0.439	0.365	0.39	-								
NMIMSM	0.56	0.414	0.585	0.609	0.56	0.585	0.365	0.585	0.536	0.439	0.341	0.463	0.341	-							
IMTG	0.536	0.414	0.609	0.56	0.487	0.585	0.414	0.536	0.536	0.439	0.341	0.487	0.39	0.512	-						
NITIEM	0.585	0.463	0.609	0.585	0.463	0.536	0.39	0.56	0.56	0.463	0.365	0.487	0.365	0.512	0.512	-					
SIBMP	0.463	0.365	0.512	0.487	0.39	0.487	0.439	0.414	0.365	0.487	0.39	0.414	0.39	0.341	0.365	0.414	-				
XIMBB	0.682	0.536	0.756	0.731	0.609	0.756	0.487	0.56	0.609	0.536	0.463	0.56	0.439	0.536	0.56	0.536	0.463	-			
TISSM	0.56	0.487	0.609	0.585	0.463	0.487	0.39	0.512	0.487	0.414	0.317	0.414	0.341	0.439	0.439	0.439	0.365	0.585	-		
IITM	0.487	0.341	0.512	0.487	0.365	0.39	0.292	0.463	0.439	0.365	0.317	0.39	0.268	0.39	0.39	0.463	0.292	0.439	0.365	-	
IITD	0.682	0.487	0.731	0.707	0.585	0.658	0.463	0.585	0.609	0.536	0.39	0.487	0.414	0.487	0.536	0.512	0.414	0.658	0.585	0.439	

Appendix D. PAD Matrix

	IIMA	IIMB	IIMC	ISBH	IIML	XLRIJ	FMSD	IIMI	IIMK	IIFTD	SPJM	MDIG	JBIMSM	NMIMSM	IMTG	NITIEM	SIBMP	XIMBB	TISSM	IITM	
IIMB	0.786	-																			
IIMC	0.933	0.781	-																		
ISBH	0.958	0.806	0.921	-																	
IIML	0.83	0.74	0.794	0.878	-																
XLRIJ	0.873	0.8	0.891	0.916	0.812	-															
FMSD	0.61	0.708	0.709	0.633	0.653	0.724	-														
IIMI	0.848	0.727	0.811	0.865	0.813	0.769	0.641	-													
IIMK	0.818	0.609	0.811	0.805	0.745	0.80	0.603	0.766	-												
IIFTD	0.709	0.705	0.738	0.730	0.727	0.721	0.857	0.75	0.678	-											
SPJM	0.654	0.636	0.620	0.642	0.666	0.629	0.714	0.612	0.571	0.80	-										
MDIG	0.723	0.72	0.718	0.741	0.74	0.80	0.75	0.763	0.727	0.745	0.772	-									
JBIMSM	0.642	0.666	0.644	0.631	0.653	0.654	0.744	0.60	0.60	0.782	0.820	0.711	-								
NMIMSM	0.741	0.666	0.738	0.793	0.727	0.786	0.612	0.857	0.785	0.692	0.622	0.745	0.608	-							
IMTG	0.709	0.666	0.769	0.73	0.727	0.786	0.693	0.785	0.785	0.692	0.622	0.784	0.695	0.807	-						
NITIEM	0.774	0.745	0.769	0.761	0.690	0.721	0.653	0.821	0.821	0.769	0.666	0.784	0.652	0.807	0.807	-					
SIBMP	0.644	0.625	0.677	0.666	0.653	0.689	0.782	0.641	0.566	0.816	0.761	0.708	0.744	0.571	0.612	0.693	-				
XIMBB	0.823	0.771	0.873	0.869	0.819	0.895	0.727	0.741	0.806	0.758	0.705	0.807	0.692	0.758	0.793	0.758	0.690	-			
TISSM	0.741	0.745	0.769	0.761	0.690	0.754	0.653	0.75	0.714	0.653	0.577	0.666	0.608	0.692	0.692	0.692	0.612	0.827	-		
IITM	0.701	0.608	0.70	0.689	0.60	0.571	0.545	0.745	0.784	0.638	0.650	0.695	0.536	0.680	0.680	0.808	0.545	0.679	0.638	-	
IITD	0.823	0.701	0.845	0.840	0.786	0.805	0.690	0.774	0.806	0.758	0.627	0.701	0.653	0.689	0.758	0.724	0.654	0.843	0.827	0.679	

VISUALIZATION OF THE SIGNIFICANT EXPLICATIVE CATEGORIES USING CATANOVA METHOD AND NON-SIMMETRICAL CORRESPONDENCE ANALYSIS FOR EVALUATION OF PASSENGER SATISFACTION

Ida CAMMINATIELLO¹

PhD, Researcher,
University "Federico II" of Naples, Italy

E-mail: camminat@unina.it

Luigi D'AMBRA²

PhD, University Professor,
University "Federico II" of Naples, Italy

E-mail: dambra@unina.it



Abstract: ANalysis Of VAriance (ANOVA) is a method to decompose the total variation of the observations into sum of variations due to different factors and the residual component. When the data are nominal, the usual approach of considering the total variation in response variable as measure of dispersion about the mean is not well defined. Light and Margolin (1971) proposed CATegorical ANalysis Of VAriance (CATANOVA), to analyze the categorical data. Onukogu (1985) extended the CATANOVA method to two-way classified nominal data. The components (sums of squares) are, however, not orthogonal. Singh (1996) developed a CATANOVA procedure that gives orthogonal sums of squares and defined test statistics and their asymptotic null distributions. In order to study which exploratory categories are influential factors for the response variable we propose to apply Non-Symmetrical Correspondence Analysis (D'Ambra and Lauro, 1989) on significant components. Finally, we illustrate the analysis numerically, with a practical example.

Key words: ANOVA; CATANOVA; Non-Symmetrical Correspondence; Passenger Satisfaction

1. Model

Many authors analyzed categorical data taking their bearings from quantitative statistics. Some of this methods require transformation of the data before analysis, others (Light and Margolin, 1971; Onukogu, 1985) do not. We start from Onukogu's approach.

Let A, B, C , be the two explicative variables and the response respectively. Let $i=1,2,\dots,I$ index the categories of C , $j=1,2,\dots,J$ index the categories of A and $k=1,2,\dots,K$ index the categories of B . Let n the number of units. Denote by N the three-way contingency table and by n_{ijk} the joint frequency. Let $n_{.jk} = \sum_{i=1}^I n_{ijk}$.

Onukogu developed the following linear model for analysis of data from three-way contingency table.

$$E(n_{ijk}/n_{.jk}) = \mu_i + \tau_{ij} + \beta_{ik} + \gamma_{ijk} \quad (1)$$

where μ_i , τ_{ij} , β_{ik} and γ_{ijk} represent the constant, j -th A effect, k -th B effect and their interaction for the i -th response, respectively.

Under model 2, the null (H_0) and alternative (H_1) hypotheses for testing the A effect, B effect and their interaction effect are defined as

$$\begin{aligned} H_{0A} : \tau_{ij} &= 0, & H_{1A} : \tau_{ij} &\neq 0 \\ H_{0B} : \beta_{ik} &= 0, & H_{1B} : \beta_{ik} &\neq 0 \\ \text{and} \\ H_{0A} : \gamma_{ijk} &= 0, & H_{1A} : \gamma_{ijk} &\neq 0 \end{aligned} \quad (2)$$

respectively.

The purpose of CATANOVA is to obtain Sums of Squares (SS) and tests for these hypotheses.

2. Sums of Squares decomposition for categorical data

One of hurdles to be cleared in any analysis of variance concerns the definition and computation of Sums of Squares (SS). When the data are nominal, the usual approach of considering the total variation in response variable as measure of dispersion about the mean is not well defined. One way out is to introduce the analysis of variance in vector notation and in terms of projectors.

Let \mathbf{A} , \mathbf{B} and \mathbf{C} be the binary indicator matrix related to the complete disjunctive coding of variables A , B , and C respectively. Let \mathbf{Z}_{AB} be the indicator matrix that represent the interaction effect between A and B . The contingency table can be constructed as

$$\mathbf{N} = \mathbf{C}^T \mathbf{Z}_{AB} \quad (3)$$

Denote by \mathfrak{R}_j the subspace generated by the columns of \mathbf{A} and \mathfrak{R}_j^\perp its orthocomplement subspace. The projection operators on \mathfrak{R}_j and \mathfrak{R}_j^\perp are constructed as

$$\mathbf{P}_A = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T, \quad \mathbf{P}_A^\perp = \mathbf{I}_N - \mathbf{P}_A \quad (4)$$

In the same way we define:

$$\begin{aligned} \mathbf{P}_B &= \mathbf{B}(\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T, & \mathbf{P}_B^\perp &= \mathbf{I}_N - \mathbf{P}_B \\ \mathbf{P}_{AB} &= \mathbf{Z}_{AB}(\mathbf{Z}_{AB}^T \mathbf{Z}_{AB})^{-1} \mathbf{Z}_{AB}^T, & \mathbf{P}_{AB}^\perp &= \mathbf{I}_N - \mathbf{P}_{AB} \\ \mathbf{P}_M &= \mathbf{1}_N(\mathbf{1}_N^T \mathbf{1}_N)^{-1} \mathbf{1}_N^T, & \mathbf{P}_M^\perp &= \mathbf{I}_N - \mathbf{P}_M \end{aligned} \quad (5)$$

The Total Sum of Squares (TSS), the Between Sum of Squares (BSS) and the Within Sum of Squares (WSS) are constructed as

$$\begin{aligned} TSS &= tr(\mathbf{C}^T \mathbf{P}_M \mathbf{C}) \\ BSS &= tr(\mathbf{C}^T (\mathbf{P}_{AB} - \mathbf{P}_M) \mathbf{C}) \\ WSS &= tr(\mathbf{C}^T \mathbf{P}_{AB} \mathbf{C}) \end{aligned} \quad (6)$$

To study the relationship between the response and the explicative variables, Light e Margolin defined the following directional measure:

$$R^2 = \frac{BSS}{TSS} \quad (7)$$

To test if the measure is significant, they proposed:

$$C_0 = (n-1)(I-1)R^2 \cong \chi^2_{(I-1)(JK-1)} \quad (8)$$

If the dependence relationship between the response and the explicative variables is significant, we proceed to test the different effects. Onukogu defined the following SS for testing different effects:

$$\begin{aligned} SS_A &= tr(\mathbf{C}^T (\mathbf{P}_A - \mathbf{P}_M) \mathbf{C}) \\ SS_B &= tr(\mathbf{C}^T (\mathbf{P}_B - \mathbf{P}_M) \mathbf{C}) \\ InteractionSS &= tr(\mathbf{C}^T (\mathbf{P}_{AB} - \mathbf{P}_A - \mathbf{P}_B + \mathbf{P}_M) \mathbf{C}) \end{aligned} \quad (9)$$

where SS_A , SS_B and $InteractionSS$ are the sum of squares due to the A effect, B effect and their interaction effect.

If there is independence between the explicative variables, TSS can be decomposed as

$$TSS = SS_A + SS_B + InteractionSS + WSS \quad (10)$$

For testing main and interaction effects, Onukogu defined the following tests

$$\begin{aligned} \chi_A^2 &= (n-1)(I-1) \frac{SS_A}{TSS} \cong \chi^2_{(J-1)} \\ \chi_B^2 &= (n-1)(I-1) \frac{SS_B}{TSS} \cong \chi^2_{(K-1)} \\ \chi_{AB}^2 &= (n-1)(I-1) \frac{InteractionSS}{TSS} \cong \chi^2_{(J-1)(K-1)} \end{aligned} \quad (11)$$

If there is association between the explicative variables, the previous components SS are not orthogonal and the decomposition (10) is not true. So Singh (1996) defined the following adjusted SS.

$$\begin{aligned}
 SS_{A/B} &= tr(\mathbf{C}^T \mathbf{P}_{A/B} \mathbf{C}) \\
 &= tr(\mathbf{C}^T (\mathbf{I}_N - \mathbf{P}_B) \mathbf{A} (\mathbf{A}^T (\mathbf{I}_N - \mathbf{P}_B) \mathbf{A})^{-1} \mathbf{A}^T (\mathbf{I}_N - \mathbf{P}_B) \mathbf{C}) \quad (12)
 \end{aligned}$$

$$\begin{aligned}
 SS_{B/A} &= tr(\mathbf{C}^T \mathbf{P}_{B/A} \mathbf{C}) \\
 &= tr(\mathbf{C}^T (\mathbf{I}_N - \mathbf{P}_A) \mathbf{B} (\mathbf{B}^T (\mathbf{I}_N - \mathbf{P}_A) \mathbf{B})^{-1} \mathbf{B}^T (\mathbf{I}_N - \mathbf{P}_A) \mathbf{C}) \quad (13)
 \end{aligned}$$

where $SS_{A/B}$ is the adjusted SS due to A variable effect after eliminating the B effect and $SS_{B/A}$ is the adjusted SS due to B variable effect after eliminating the A effect.

The interaction SS is obtained by subtraction as

$$\begin{aligned}
 IntSS &= BSS - SS_A - SS_{A/B} \\
 &= BSS - SS_B - SS_{B/A} \quad (14)
 \end{aligned}$$

As consequence TSS can be decomposed as

$$\begin{aligned}
 TSS &= SS_A + SS_{B/A} + IntSS + WSS \\
 &= SS_B + SS_{A/B} + IntSS + WSS \quad (15)
 \end{aligned}$$

For testing main and interaction effects, Singh (1996) defined the following tests

$$\begin{aligned}
 C_{01} &= (n-1)(I-1) \frac{SS_{A/B}}{TSS} \cong \chi^2_{(I-1)(J-1)} \\
 C_{02} &= (n-1)(I-1) \frac{SS_{B/A}}{TSS} \cong \chi^2_{(I-1)(K-1)} \\
 C_{012} &= (n-1)(I-1) \frac{IntSS}{TSS} \cong \chi^2_{(I-1)(J-1)(K-1)} \quad (16)
 \end{aligned}$$

3. Analysis of significant components

CATANOVA enables us to know if there is significant dependence between independent and dependent variables and which exploratory variables are significant to explain the response, but which exploratory categories are influential for the response variable?

In order to describe the dependence relationship between independent and dependent variables we propose to carry out Non-Symmetrical Correspondence Analysis (NSCA).

NSCA looks for the orthonormal basis which accounts for the largest part of inertia to visualize the dependence structure between the variables in a lower dimensional space. This leads us to the extraction of the eigenvalues λ_α and eigenvectors \mathbf{u}_α associated to the eigen-system

$$\left(\frac{1}{n}\right)\mathbf{C}^T(\mathbf{P}_{AB} - \mathbf{P}_M)\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \quad (17)$$

NSCA enables us to measure and visualize the strength of the asymmetrical relationship between the dependent and independent categories, to analyze which categories are significant for the response, to carry out confidence circles for identifying those categories that are not statistically influential in helping to explain the response.

The focus of this paper is to explore the significant components by NSCA.

Let us consider the matrix of the A variable effect after eliminating the B effect

$$\mathbf{S}_{A/B} = \underbrace{\mathbf{C}^T(\mathbf{I}_N - \mathbf{P}_B)}_{\mathbf{Q}^T} \underbrace{\mathbf{A}(\mathbf{A}^T(\mathbf{I}_N - \mathbf{P}_B)\mathbf{A})^{-1}\mathbf{A}^T}_{\mathbf{D}^-} \underbrace{(\mathbf{I}_N - \mathbf{P}_B)\mathbf{C}}_{\mathbf{Q}} \quad (18)$$

where \mathbf{D}^- is a generalized inverse. NSCA leads us to the extraction of the eigenvalues λ_α and eigenvectors \mathbf{u}_α associated to the eigen-system

$$\mathbf{Q}^T \mathbf{D}^- \mathbf{Q} \mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \quad (19)$$

The response and A variable coordinates on α axis are given by

$$\boldsymbol{\psi}_\alpha = \sqrt{\lambda_\alpha} \mathbf{u}_\alpha \quad \boldsymbol{\varphi}_\alpha = \mathbf{D}^{-1/2} \mathbf{Q} \mathbf{u}_\alpha \quad (20)$$

respectively. It is easy to verify that

$$\boldsymbol{\varphi}_\alpha^T \boldsymbol{\varphi}_\alpha = \mathbf{u}_\alpha^T \mathbf{Q}^T \mathbf{D}^- \mathbf{Q} \mathbf{u}_\alpha = \lambda_\alpha \quad \sum_\alpha \boldsymbol{\varphi}_\alpha^T \boldsymbol{\varphi}_\alpha = \sum_\alpha \lambda_\alpha \quad (21)$$

These coordinates are especially useful for describing the dependence relationship between the dependent and independent variables. In particular, A variable coordinates close to the origin will infer that their categories do not help predict the response categories. Response coordinates close to the origin indicate that very few explanatory categories are influential in determining the outcome of those response categories. Similarly coordinates far from the origin will highlight that, if they are associated with the explanatory variable, those categories are influential factors for the response variable. If a response category lies far from the origin then there will be explanatory factors that influence its position.

To complement the correspondence plots, more formal tests of the influence of particular categories may be made by considering the confidence circles for NSCA proposed by Beh and D'Ambra (2005). If one considers only the dependence between the row

(predictor) and column (explanatory) variable, the C-statistic can be expressed in terms of the predictor coordinates such that

$$C_0 = \frac{(n-1)(I-1)}{\left(1 - \sum_{i=1}^I p_{i..}^2\right)} \sum_{j=1}^J \sum_{\alpha=1}^M p_{j\alpha} \phi_{j\alpha}^2 \cong \chi_{(I-1)(J-1)}^2 \quad (22)$$

where $p_{i..}$ and $p_{.j}$ are the i -th and j -th marginal proportions so that $\sum_{i=1}^I p_{i..} = \sum_{j=1}^J p_{.j} = 1$. Beh and D'Ambra showed that 95% confidence circles for the j explanatory column category represented in a two dimensional non-symmetrical correspondence plot has radii length

$$r_j^J = \sqrt{\frac{5.99 \left(1 - \sum_{i=1}^I p_{i..}^2\right)}{p_{.j} (n-1)(I-1)}} \quad (23)$$

Note that (23) depends on the j -th marginal proportion. Thus, for a very small classification in the j -th(explanatory) category, the radii length will be relatively large. Similarly, for a relatively large classification, the radii length will be relatively small.

4. A numerical example

The data was collected in a enterprize of Local Public transport in Naples. We will treat Satisfaction as the response variable and age and profession as the predictor variables. The response is measured on a scale ranging from 1 (Low) to 4 (High), the age has five categories (< 18, 19-25, 26-40, 41-65, > 65), also the profession has five categories (student, employee, housewife, pensioner, other). The passengers are 400.

For our table $R^2 = 0.06$ which has an associated C_0 -statistic of 73.61 (p-value =0.001). Therefore we can conclude that the age and profession influence the Passenger Satisfaction. To further investigate the source of this asymmetrical relationship, we carry out the CATANOVA.

Table 1 shows the results of decomposition of $TSS = SS_A + SS_{B/A} + IntSS + WSS$.

Table 1. Decomposition of $TSS = SS_A + SS_{B/A} + IntSS + WSS$

Source	d.f.	SS	Test	p-value
Age (adjusted)	12	4.857	23.161	0.008
Profession	4	3.447	16.439	0.001
Age x Profession	48	7.134	34.016	0.016
Within	335	235.595		
Total	399	251.035		

Table 2 shows the results of decomposition of $TSS = SS_B + SS_{A/B} + IntSS + WSS$.

Table 2. Decomposition of $TSS = SS_B + SS_{A/B} + IntSS + WSS$

Source	d.f.	SS	Test	p-value
Profession (adjusted)	12	3.060	14.595	0.0583
Age	4	5.244	25.005	0
Age x Profession	48	7.134	34.016	0.016
Within	335	235.595		
Total	399	251.035		

CATANOVA shows that the age is a more influential factor in level of satisfaction than the profession. Moreover the age effect after eliminating the profession effect is significant at 1 %, the profession effect after eliminating the age effect is significant at 6 %. So we proceed to apply the NSCA on the matrix related age effect after eliminating the profession effect. Of course we could carry out the NSCA of each component.

The results of NSCA show that the first two factors explain the 97 % of variability, so they allow us to have a quite complete view of the dispersion of phenomenon.

The plan produced by first two factors (Figure 1) shows that passengers who are less than 18 and between 41 to 65 tend to be not too satisfied, those between 18 to 25 tend to be pretty satisfied, those who are more than 65 and between 26 to 40 tend to end up either not satisfied or very satisfied.

For Figure 1 95 % confidence circles have been included. Since the origin, which is associated with zero predictability of the response variable given the explanatory variables (ie. independence), does not lie within any of the circles, all of the categories of the age variable are statistically influential in helping to determine the passenger satisfaction.

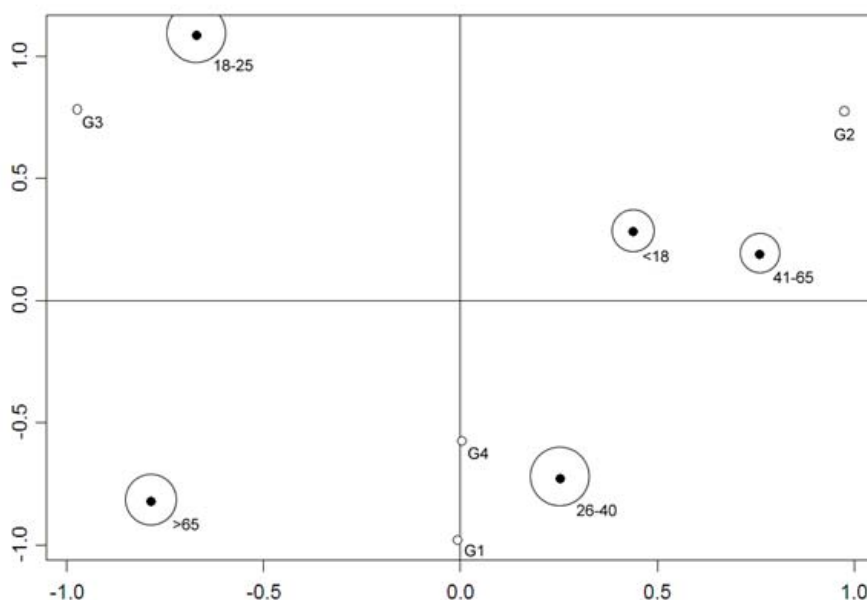


Figure 1. NSCA plot

5. Final remarks

We applied CATANOVA to study if there is significant association between independent and dependent variables and which exploratory variables are significant to explain the response.

In order to measure and visualize the strength of the asymmetrical relationship between the dependent and independent variables and to study which categories are significant to explain the response we proposed to carry out the NSCA of significant factors.

The combined approach has been applied on the data collected in a enterprise of Local Public transport in Naples obtaining interesting results

The paper focused on nominal data. In the next we will focus on ordered categorical variables.

References

1. Anderson, R. G. and Landis, R. **Catanova for multidimensional contingency tables: nominal-scale response**, Communications in Statistics, 9, 1980, pp. 1191-1206
2. D'ambra, A., Ciavolino, E. and De Franco, D. **Un approccio statistico multivariato per la valutazione dell'utente nell'ambito dei trasporti: il caso AMTS**, "Proceedings of MTISD'06 Conference", Napoli, 2006
3. D'ambra, L., Beh, E. J. and Amenta, P. **Catanova for two-way contingency tables with ordinal variables using orthogonal polynomials**, Communications in Statistics, Theory and Methods, 34, 2005, pp. 1755-1769
4. D'ambra, L. and Lauro, N. C. **Non symmetrical analysis of three-way contingency tables**, In Coppi, R. and Bolasco, S. (Eds.), Multiway Data Analysis, North Holland, 1989, pp. 301-314
5. Light, R. and Margolin, B. **An analysis of variance for categorical data**, Journal of the American Statistical Association, 66, 1971, pp. 534-544
6. Onukogu, I. B. **An analysis of variance of nominal data**, Biometrics Journal, 27, 1985, pp. 375-384
7. Singh, B. **On CATANOVA method for analysis of two-way classified nominal data**, Sankhya: The Indian Journal of Statistics, 58, 1996, pp. 379-388
8. Takeuchi, K., Yanai, H. and Mukeherjee, B. N. **The foundations of multivariate analysis**, Wiley, New York, 1981

¹ Ida Camminatiello is currently contract researcher (Formez). She holds a PhD in statistics from the University of Naples Federico II. She worked as a Lecturer at the Faculty of Economics, Second University of Naples, where she taught statistics, time series analysis and informatics. She has participated to numerous national and international conference. The main research topics are related to robust regression, multinomial logit model, categorical analysis of variance and non-symmetrical correspondence analysis with applications in environmental field, transport, customer satisfaction and sensory analysis. The most important and recent publications are:

1. Camminatiello, I., D'Ambra, L., Meccariello, G. and Della Ragione, L. **A study of instantaneous emissions through the decomposition of directional measures for three-way contingency tables with ordered categories**, Journal of Applied Sciences (to appear).
2. Camminatiello, I. and Lucadamo, A. **Estimating multinomial logit model with multicollinear data**, Asian Journal of Mathematics and Statistics (to appear).
3. Lombardo, R. and Camminatiello I. **CATANOVA for two-way cross classified categorical data**, Statistics: A Journal of Theoretical and Applied Statistics, 2009
4. Camminatiello, I. **A robust approach for partial least squares regression**, in "Metodi, modelli e tecnologie dell'informazioni a supporto delle decisioni, II. Applicazioni" Franco Angeli, Milano, 2008, pp. 31-38,

5. Camminatiello, I. and D'Ambra A. **Evaluation of Passenger Satisfaction using three-way contingency table with ordinal variables**, *Rivista di Economia e Statistica del Territorio*, 1, 2008, pp. 25-38

² Luigi D'Ambra is a professor of statistics, University of Naples Federico II. In December 2005 he has been invited by a School of Quantitative Methods and Mathematical Sciences (University of Western Sydney) to participate to several conferences and to carry out research activity (Project title: The non-symmetrical correspondence analysis of ordinal categorical data). He has participated to numerous national and international conferences as invited lecturer and discussant. He has been the chairman of steering committee of the schools of Italian Statistics Society on the "Statistical Methods for Customer Satisfaction Evaluation" and "Statistical Methods for the Healthcare Services Evaluation". He has been School President of the European Courses in Advanced Statistics - ECAS 2003 too. The main research topics are related to the non-symmetric techniques, to the analysis of multi-way data tables, with respect to qualitative and quantitative variables, and to the interpretative aspects of automatic classification. The most important and recent publications are:

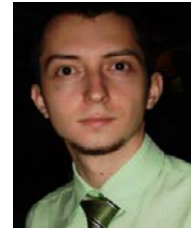
1. Beh, E. and D'ambra, L. **Some interpretative tools for non-symmetrical correspondence analysis**, *Journal of Classification*, vol. 26, 2009, pp. 55-76
2. Beh, E., Simonetti, B. and D'ambra, L. **Partitioning a non-symmetric measure of association for three-way contingency tables**, *Journal of Multivariate Analysis*, vol. 98, 2007, pp. 1391-1411
3. Lombardo, R., Beh, E. and D'ambra, L. **Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials**, *Computational Statistics & Data Analysis*, vol. 52, 2007, pp. 566-577
4. Beh, E., Simonetti, B. and D'ambra, L. **Three-way ordinal non symmetrical correspondence analysis for the evaluation of the patient satisfaction**, *Statistica & Applicazioni*, 2006

IMPLEMENTING A GIS APPLICATION FOR NETWORK MANAGEMENT¹

Alexandru SMEUREANU²

PhD Candidate, Economic Informatics Department,
University of Economics, Bucharest, Romania

E-mail: alexandru.smeureanu@gmail.com



Stefan Daniel DUMITRESCU³

PhD Candidate, Computers Department,
Politehnica University of Bucharest, Bucharest, Romania

E-mail: dumitrescu.stefan@gmail.com



Abstract: *The paper proposes a network management system architecture based on a geographical information system that allows accurate description and inventorying of the infrastructure. The system contains several models that emulate real life operational networks based on fiber optics, copper and WiFi technologies. The architecture is implemented in a network management systems application and a number of interesting performance and design problems encountered during the implementation are presented along with their solutions.*

Key words: *GIS; network management system; vector/raster model; rendering performance*

1. Introduction

The evolution of the Internet in recent years has fundamentally changed the way people interact and communicate. This growth of the Internet has led to chaotic development of the infrastructure that supports it.

The development did not take account of hardware used, the interconnection media, software employed, the size of networks that are interconnected or any structured expansion plan. Most extensions were made incrementally, within the limits of the available budgets.

Difficulties arose concerning network design and description, infrastructure expansion and troubleshooting failures. Areas where the focus is on real-time data transfer (medicine, banking, police, army) are seriously affected by lack of network reliability.

Another result of the expansion of the network infrastructure is the increase in the spatial dimensions that can complicate the troubleshooting procedures done in the field by technicians if the location of equipments is not accurately noted. The best way for modeling the geographical dimension of networks is to use a geographical information system (GIS).

2. GIS system definition and concepts

A GIS integrates hardware elements, software and data for capturing, managing, analyzing, and displaying geographically related information. This system allows viewing, understanding and querying data in multiple ways that reveal relationships and patterns in the form of maps, reports or charts.

A GIS helps with answers to questions and solving problems by looking at existing data in an intuitive and easily distributed way.

A GIS can be seen in three different ways: in terms of a database (database view), the map (map view) and model (model view). In Database View the GIS is seen as a structured database that describes the world in geographical terms. In View map the GIS is seen as a set of intelligent maps and sketches which characterized relations over the Earth. In Model View the GIS system is seen as a set of tools for information transforming for obtaining new derived datasets from existing data sets. These tools extract information from existing data, apply analytic functions, and write results into new derived datasets [9]⁴.

Data representation in a GIS system can be done either in Raster or Vector modes.

Raster mode is essentially any type of digital image represented as an array of pixels. The pixel is the smallest unit of an image. A combination of these pixels will create an image. This representation consists of rows and columns of cells, each cell with one stored value. Raster data can be images (raster image) with each pixel containing a value, usually a color. Additional values recorded for each cell may be a discrete, defined by the user with relevant GIS system, a continuous value, such as temperature, or a null value if no data available. A raster cell that stores only one value can be extended by using raster bands to represent RGB colors (red, green, blue), or an extended attribute table with one row for each single cell unique value. Resolution in raster mode is pixel size in physical units (e.g. distance). Raster data are stored in various formats, from standard file structure of TIF, JPEG, etc., and directly in binary data fields (BLOB) of common databases.

Vector mode is used especially in GIS systems where geographical features are often expressed as vectors, by considering the elements as geometric forms. Various geographical elements are expressed by different types of geometry forms:

- Points - zero-dimensional points are used for geographical elements that can be best expressed by a single point of reference, in other words, simple location. There is no possibility to make any measurements in this case.
- Lines - The lines are used for one-dimensional linear elements such as rivers, roads, topographic lines, and so on. Lines also allow measurement of the distance.
- Polygons - Polygons are two-dimensional elements that are used to represent a geographical area on the surface of the Earth. Polygon features make it possible to measure the perimeter and area.

In order for the GIS to be useful, they must work properly and provide the requested information in a timely manner. The common problems of such systems are scalability and speed of processing user requests. When the system operates with more elements, the scalability problem becomes more pregnant. Problems in repeated rendering of large number of elements appear (e.g. moving the map in a specific direction) when finding a particular item, when querying the system for the position of an element, inserting or deleting an item, and so on [10].

GIS systems used for network modeling

The application integrates a GIS engine based on vector graphics. This means that the application stores all its information in the form of vector primitives: points, lines and polygons. This enables the GIS rendering system to manipulate the information as a vector image.

The geographical component of the data describing the network is divided into two major layers. Each of the layers is later subdivided into several sub layers.

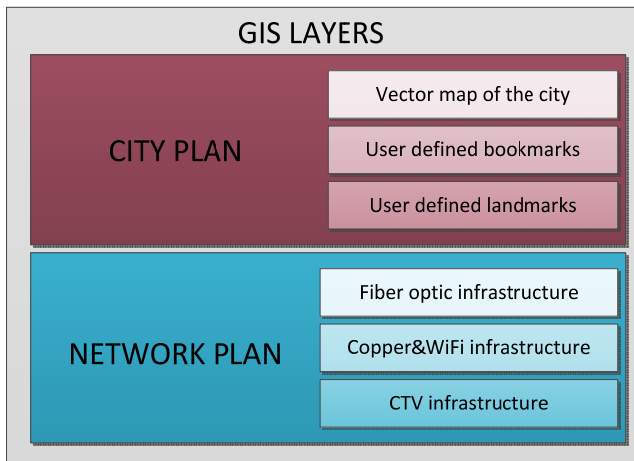


Figure 1. GIS Layers

3. The City Plan

The first major layer is the City Plan that contains information about the terrain topology on which the network operates. The City Plan includes sub layers that describe streets, buildings, duct systems, user defined landmarks. The user defined landmarks are geographic location markers that hold a particular importance for user.

Because a plan of a city contains large amounts of information that is not directly important and is rarely updated, the layer is stored primarily in a vector graphics image format file that is downloaded and stored on the user computer. This is done in order to minimize the resources usage on the database servers because having the city plan stored locally prevents the transfer of redundant information over the network (considering that the city plan is rarely updated) and also saves the server from running exhaustive spatial queries.

The file format used for storing the vector image of the city is Enhanced Metafile - EMF which is a file format type developed by Microsoft intended to be a portable file format between applications. EMF allows the storage and manipulation of both vector graphics and raster graphics in the same file and this in turn enables the creation on mixed maps containing both vector images (where the map vectorization was possible) and also raster image regions (such as satellite images).

There are also downsides to the fact that the vector map is saved locally, being that EMF is a file storage format that offers no spatial indexing or other type of indexing on top of a space filling algorithm (almost all file formats do not have this feature). The lack of any indexing method affects the application's performance. Rendering a rectangle region from vector map with no index can take around two seconds for a 66MB map containing over two million objects.

A quick analysis of the type of queries that need to be run by the GIS engine will show that they involve in most cases the retrieval of a rectangle-shaped region. The height and width of the region are almost all the time proportional to the application window size and to the zoom level selected by the user.

In order to improve the performance of the GIS map rendering subsystem, several practical improvements can be made: splitting the map into sectors and raster caching some of the queries.

There are two feasible approaches when working with large vector images containing large maps:

- **Method 1** - Raster caching important zoom levels and splitting the image in sectors. The storage is done in raster format indexed on 2D coordinates. Every time a region part of a cached zoom level is requested it is identified and retrieved very fast. Each time the query requests a region that is not cached the system uses the vector map to generate it.
- **Method 2** - Splitting the actual image into sectors and storing them indexed on 2D coordinates. When a region is requested the sectors that contain parts of it are queried individually and the result is formulated.

In order to evaluate the performance of the methods mentioned above several testing scenarios were developed and ran:

- **Scenario 1** - Map of Bucharest in EMF version, raster caching important zoom levels (**Method 1**), random regions from the map are requested corresponding to different levels of zoom (random walk on the map).
- **Scenario 2** - Map of Bucharest in EMF version, vector splitting of the map (**Method 2**), random regions from the map are requested corresponding to different levels of zoom (random walk on the map).
- **Scenario 3** - Map of Bucharest in EMF version, raster caching important zoom levels (**Method 1**), test case containing access queries for regions corresponding to actual real life network.
- **Scenario 4** - Map of Bucharest in EMF version, vector splitting of the map (**Method 2**), test case containing access queries for regions corresponding to actual real life network.

The first step in implementing the performance evaluation is to define the needed key performance indicators. These indicators reflect several important aspects like resource consumption and user experience. The main resources used are hard-disk space required to store the map, RAM size needed to store all the data, processor time. Because the user interface is the same in all scenarios and all of them were tested automatically using specific test cases, it is impossible to quantify the user experience. That is why we decided to use as an indicator for user experience the responsiveness of the GIS when changing location, measured as the time between issuing a go-to a location command (jump / zoom / pan) and the time the application fully completes this command.

$$P = Hdd * C_1 + Ram * C_2 + Cpu * C_3 + 2^{T * C_4 - 1} * C_3$$

P - Performance loss indicator

Hdd - Hard drive space required to store the vector map including raster caches measured in megabytes (MB)

C_1 - Transformation constant of performance loss / MB of hard drive space used

Ram - Medium amount of random access memory used by the application during the process

C_2 - Transformation constant of performance loss / MB of RAM used

Cpu - Processor time used by the application for the entire the test case

C_3 - Transformation constant of performance loss / time in second of processor used

T - Average time in second spent in query

C_4 - Transformation constant of performance loss / average time in second of a location query

C_5 - Transformation constant of performance loss / average time in second of a location query exponential effect

The performance loss is a score based indicator that is designed to allow the comparison of the methods presented. A higher score indicates a higher loss of performance. All the resource consumption components are linearly weighted with the use of several constants. The user experience measured as application response time has an exponential effect on the performance loss indicator. By analyzing the behavior of several users we notice that they tend to get annoyed if the application takes more than one second to process a location query, being perceived as application sluggishness. The influence of the *T* - Average time indicator becomes exponentially more noticeable as queries take longer than one second. The constants $C_1..C_5$ have values proportional to the cost of the specific resource on the market: $C_1 = 10$, $C_2 = 100$, $C_3 = 100$, $C_4 = 1$, $C_5 = 1000$.

Scenario 1 and Scenario 3

Several sectors of the city in 2000 pixel by 2000 pixel format stored as jpeg files corresponding to different levels of zoom have been pre rendered from the original vector map. There two test cases one containing ten levels of zoom stored on 476 MB and the second one contains seventeen levels of zoom stored on a 1673 MB. The space required to store the map for one zoom level depends on the surface of the city map. The surface of the city is proportional to the second power of zoom level. This means that the space required increases exponentially.

Table 1. Scenario 1 and 3 results

Scenario 1	<i>Hdd</i>	<i>Ram</i>	<i>Cpu</i>	<i>T</i>	<i>P</i>
Case 1	476 MB	100 MB	10 sec	0.5 sec	16467.11
Case 2	1673 MB	100 MB	10 sec	0.2 sec	28304.35
Scenario 3	<i>Hdd</i>	<i>Ram</i>	<i>Cpu</i>	<i>T</i>	<i>P</i>
Case 3	476 MB	100 MB	15 sec	0.9 sec	17193.03
Case 4	1673 MB	100 MB	13 sec	0.5 sec	28737.11

The main differences between Case 1 and Case 2 are related to the hard disk space and average time needed solve a query. Case 2 has more levels of zoom cached and is able to render faster all the queries that are not covered by the cache used in Case 1.

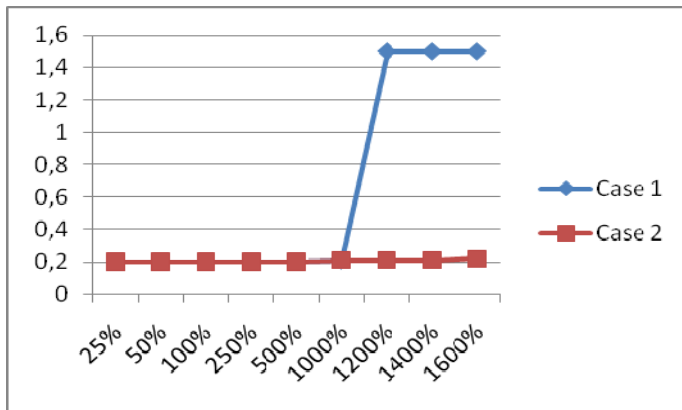


Figure 1. Average time in seconds spent in query for different zoom levels for Scenario 1

The difference between Case 3-4 and Case 1-2 can be explained by the fact that in real life users work on a small area of the map at zoom levels larger than 800% that allow them to see details about the buildings. The random walk uses a normal distribution that evenly distributes the queries on all zoom levels.

Scenario 2 and Scenario 4

The original vector map is split in several rectangular regions stored in vector format EMF file. The original vector map is no longer needed, unlike Scenario 1, as all the data is contained in the split regions.

Table 2. Scenario 2 and 4 results

Scenario 2	<i>Hidd</i>	<i>Ram</i>	<i>Cpu</i>	<i>T</i>	<i>P</i>
Case 1	50 MB	100 MB	20 sec	0.7 sec	13370.55
Scenario 4	<i>Hidd</i>	<i>Ram</i>	<i>Cpu</i>	<i>T</i>	<i>P</i>
Case 2	50 MB	100 MB	20 sec	0.7 sec	13370.55

Because in Method 2 (Scenario 2 and 4) the data is rendered from a vector source, the CPU is used for a longer period of time than in Method 1 (Scenario 1 and 3) where the onscreen rendering consists in a jpeg decompression and bit copying. The average time in seconds spent in query is almost constant over all the domain with the exception of zoom levels that overview the map. This is due to the fact that overviews contain information stored in several vector regions.

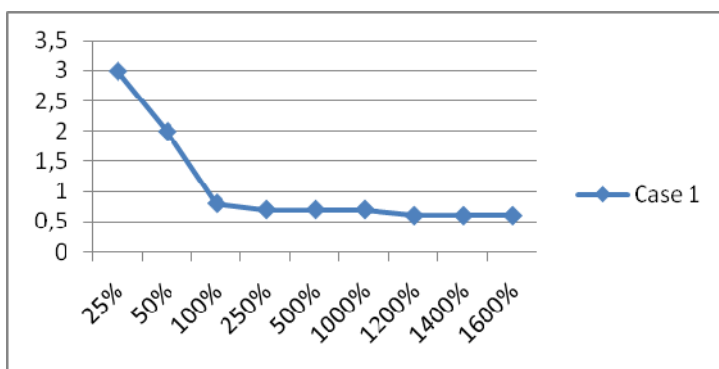


Figure 2. Average time in second spent in query for different zoom levels for Scenario 2

When using a real life data set (Case 2) we don't notice any change in the query time (vs. Case 1). The response time is similar because at every zoom level the time needed depends only on the number of regions needed to render the response.

Both approaches have their own strengths and weakness. The raster caching of entire zoom levels offers great performance as long as you are within the limits of a zoom level that is cached and poor outside these limits. At zoom levels that overview the map the disk space required to store the pre rastered sections of the map is low but a closer look zoom level will require a much larger disk footprint. The growth of the file size is the second power exponential of the multiplication index. Going outside the cached zoom levels the rendering time greatly increases. This method has great performance when working with low multiplication indexes.

The splitting of the vector image in several vector sectors has its advantages when working with high multiplication indexes. The number of vector sectors needed in rendering a region decreases with multiplication indexes because the probability of the region to be found on the border of two or more regions decreases with the relative difference between the size of the sector and the size of the requested region. The performance of the method is directly linked to the number of sectors used. In overview zoom levels the method performs poorly as it needs to query several sectors that are caught within the field of view. Zoom levels with high multiplication indexes perform very well because they can be rendered with a low number of sectors. This method can be improved to work well if a raster cache is used to store results for later use in situations where the number of sectors involved in rendering is larger than a defined value.

By using one of the two method mentioned above the downside of having no spatial index to improve searching in the EMF file can be overcome within acceptable performance limits.

4. The Network Plan

The Network Plan contains all the important information about the description of the network. The layer is structured in several sub-layers each of them embedding a model for describing and monitoring the infrastructure based on specific transmission mediums.

The rendering system for the Network Plan is based on vector graphics. Because the application has to be used in troubleshooting the physical network problems it has to be able to run in insolated mode (without requiring any network access). As previously discussed, the City Plan is already stored locally and rarely updated (the only time a network connection is needed) this meaning that the Network Plan also has to work in a limited offline mode.

Another issue investigated in [8] reveals a scalability problem when applications are working with spatial databases in remote locations. This reveals the need to implement a spatial indexing algorithm in the local application in order to cope with large number of queries (e.g. like the one generated by moving mouse over different objects and indentifying which of them is hovered over by the cursor). Implementing a spatial indexing algorithm make sense only if the data which is indexed is accessible locally, either in RAM or worst case on the hard-disk.

These two issues sustain the idea of creating a local data storage system that enables part of the data that is vital to be stored locally. The data will be synchronized with

the remote database only when changes appear either locally or on the remote side in the database as a result of another client working with the application. Having the network description data locally enables the application to work desynchronized in restricted mode when the network connection becomes unavailable. The synchronization uses a versioning mechanism that stores the timestamp of the latest change at data row level. The timestamp is always assigned based on the database server's current time and the process is done by using data triggers that update the information stored in the version field. The database also keeps the timestamp associated with the last change performed in the data table.

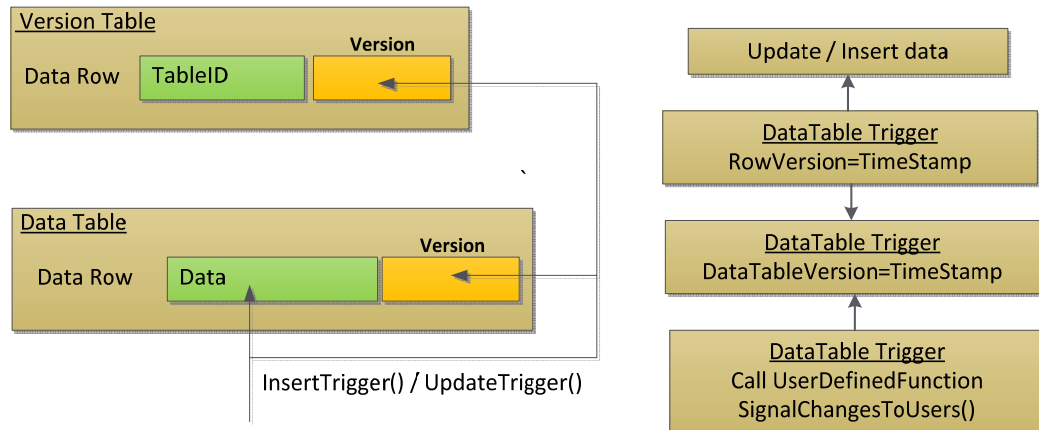


Figure 3. Database schema example

The synchronization process is very fast. The application checks the Version Table and asks which data table has a version number greater than the version in the local storage. If such a data table is detected, a query is immediately run on the data table to retrieve all entries that have a higher version number. After retrieving the differences, the local database is synchronized: the new entries are added and the old ones are updated/removed.

5. Network management system implementation

The application allows the accurate description of physical network equipments by modeling key technology elements that compose the network infrastructure. The modeled technologies are fiber optics, copper cable infrastructure based on UTP twisted pairs, WiFi point to point links and analogue CTV coaxial infrastructure. Each technology can be used for data transmission and has its own specific equipments.

Fiber optic modeling

Fiber optics infrastructure can be composed out of several major components: fiber optics cable, optical distribution frames, enclosures, media convertors, passive optical multiplexers and demultiplexers.

Modeling Fiber Optic Cabling

Fiber Optic is used as a medium for telecommunication and networking because it is flexible and can be bundled as cables. Fiber-optic cabling is made either of glass or plastic and used to guide light impulses. It is especially advantageous for long-distance

communications, because light propagates through the fiber with little attenuation compared to electrical cables. This allows long distances to be spanned with few repeaters.

Current transmission standards have yet to approach the physical potential bandwidth of this medium.

Generally, an optic cable contains multiple buffers that in turn contain several threads that are the actual carriers of light. Fiber optic cables can be broadly classified into two types, based on the source that emits the light: single-mode and multimode. Single-mode optical fiber carries a single ray of light, usually emitted from a laser. Because the laser light is unidirectional and travels down the center of the fiber, this type of fiber can transmit optical pulses for very long distances. Multimode fiber typically uses LED emitters that do not create a single coherent light wave. Instead, light from a LED diode enters the multimode fiber at different angles.

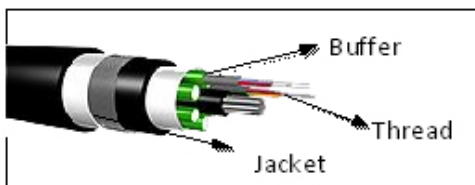


Figure 4. Fiber optic structure

In order to convert light impulses to electric signals special equipments named media convertors are used. In order to communicate, two hosts need at least two wires in order to achieve full duplex communication: one for TX transmission and one for receiving RX. Dual fiber media convertors use two threads, one for transferring RX and one for TX. Media convertors that use Wavelength Division Multiplexing (WDM) technology use two or more different wavelengths for transferring light impulses; by doing this they can achieve multiple connections over the same fiber thread.

The fiber optic model is able to store information about: geographical coordinate of cable positioning, cable reserves present on different point of the fiber, information about the type of the fiber, number of buffers, number of threads and the connections at each end of the thread.

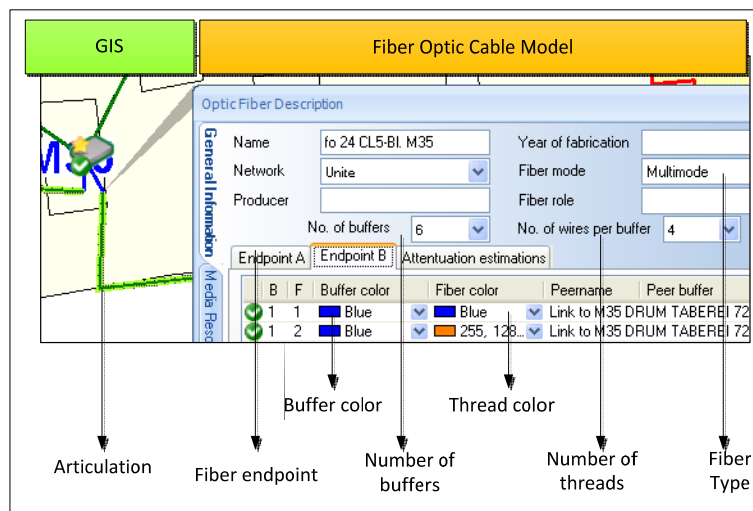


Figure 5. Fiber optic cable model

Modeling an ODF - Optical distribution frames

An optical distribution frame is a fiber optic management unit used to organize the fiber optic cable connections. It is usually used indoor and the can take any size, from small, like a patch panel, to big frames. Linking two optical cables normally is done by using one of two splicing techniques: fusion splicing and mechanical splicing. Fusion splicing works by generating a high voltage electric arc that welds two fiber threads together. Mechanical splicing works by bringing two fiber threads heads close together in a gel that has a refraction index similar to the one of the optic cable allowing light to pass thought with a limited attenuation of the signal.

An optical distribution frame has two important parts: the inside splicing diskette and the pigtails that are factory connected to the outside ports and the outside port panel.

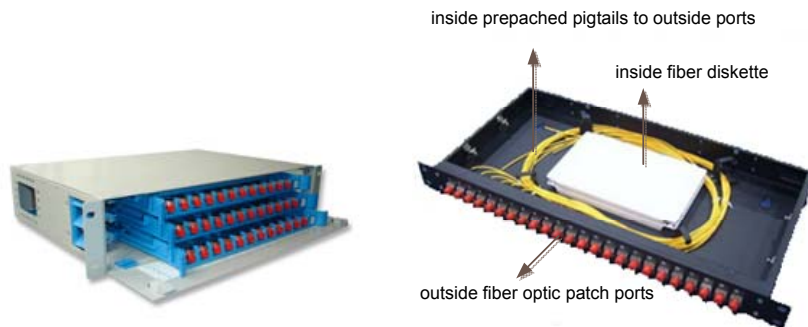


Figure 6. Figure Optical distribution frame structure

Each pigtail is welded to a fiber thread.

The benefits of using an optical distribution frame are important. First, when physically rerouting a circuit, an engineer will change only the position of a patch cable. If there is no ODF present the engineer would need an optical splicing machine. Second, the ODF reduces de complexity of the wiring - the technician will not need to know the colors of the buffer and thread in the fiber to identify a portion of the circuit. All he needs to know is the port in the ODF.

In order to describe an ODF in a network monitoring system, it has to show its two main components: the inside pigtails that are welded directly to threads from a fiber and the outside optic ports where fiber patches can be connected.

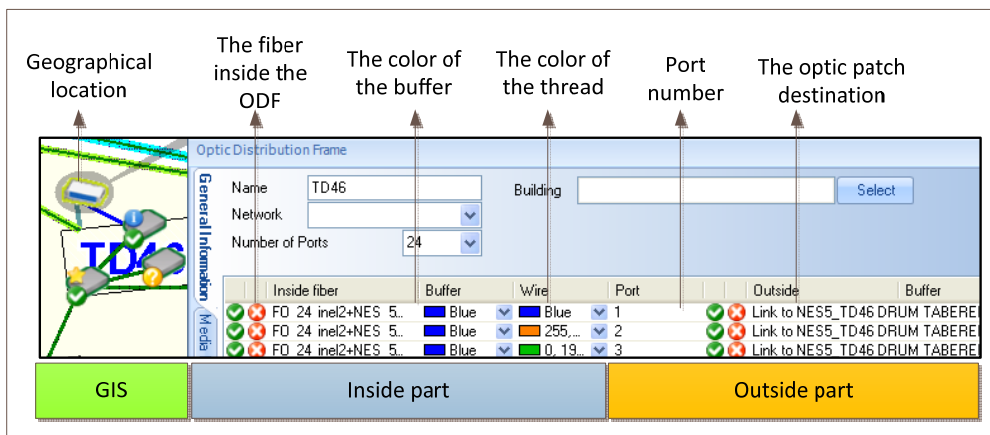


Figure 7. Optical distribution frame representation

Modeling Enclosures

The Enclosure is a fiber optic management unit used for protecting splicing from different fibers. The Enclosure has on its bottom several sockets used for incoming fiber cables that are spliced inside. On the inside, the Enclosure has several splicing diskettes that hold the welded fibers tight in place that would be otherwise vulnerable to mechanical shocks.



Figure 8. Enclosure

The Enclosure is modeled as a vector of one on one weld entries. Each weld contains two joined threads that are exactly identified by a fiber unique id and name, buffer color and thread color and a diskette number that identifies the diskette that contains the splices.

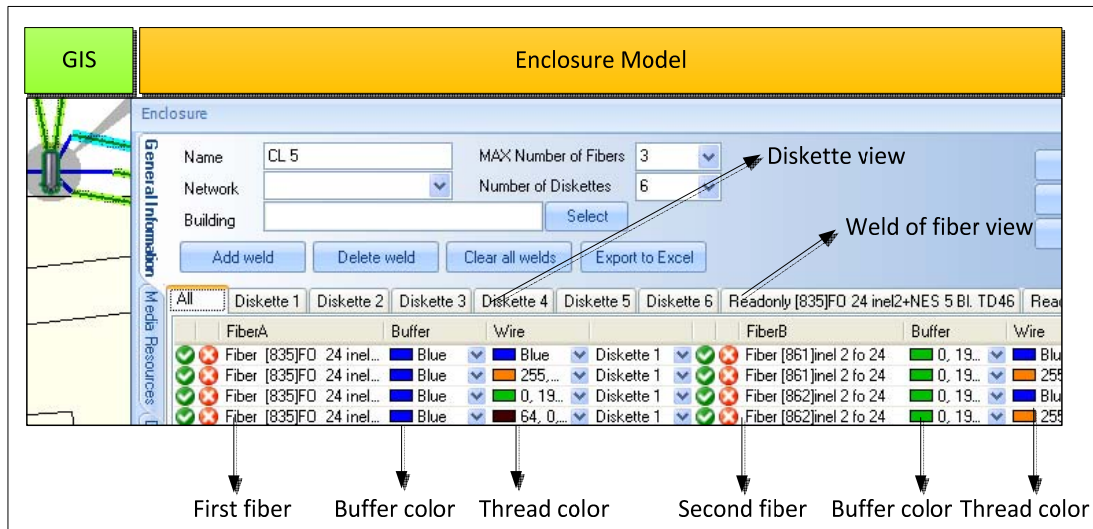


Figure 9. Optical enclosure representation

Media convertors

A media converter is a device that is used to change electrical signals on a copper cable to optical impulses that run on fiber, letting a company introduce fiber in the network without making other changes. Because of their capability, media convertors are used by networks that are in the process of upgrading from copper to fiber, and are unable to do it all at once. They have to do it incrementally either because the network is in production or they don't have the budget, manpower or the time needed.

Media converters are also used for connecting the last-mile copper connection to the optical metropolitan-area networks. The vast majority of the personal desktop computers and laptops do not have a built-in optical network card, and rely instead on the current Ethernet standard working on UTP twisted pair copper wires.

Media converters work on the physical layer of the network in accordance to the OSI stack; they do not interfere with upper-level protocol information. This lets them preserve quality of service and Layer 3 switching. They receive data signals from one media and convert them to another while remaining invisible to data traffic and other net devices. They act like bridges connecting two different communication mediums without changing the nature of the network.

The simplest form, a media converter is a small device with two media-dependent interfaces and a power supply. It can be installed almost anywhere in a network. The style of connector depends on the selection of media to be converted by the unit. In a Fast Ethernet environment, a 100Base-TX to 100Base-FX Media Converter connects a 100Base-TX twisted-pair device to a 100Base-FX compliant single or multimode fiber port that has a fiber-optic connector. In Gigabit Ethernet, a media converter is commonly deployed to convert multimode to single-mode fiber. The number of optical connectors used for fibers also varies depending on the technology used. A WDM media converter usually uses only one fiber. Media converter shelters can be used in places in which media converters are packed in large numbers. These shelters are chassis-style devices designed to be rack - mounted that can be managed with SNMP.

The media convertors are modeled as network devices that contain two ports. One of the two ports is an electrical Ethernet type port such as 1000BASE-T, 1000BASE-CX, 100BASE-TX, while the other one is an optical port such as: 1000BASE-SX multi-mode fiber that is rarely used anymore, 1000BASE-LX single-mode fiber, 10GBASE-LR or 10GBASE-ER.

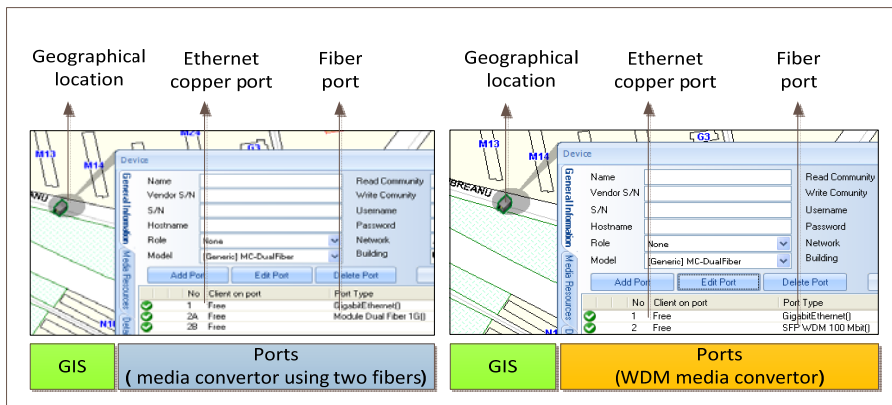


Figure 10. Device characteristics

Fiber attenuation model

The attenuation represents the extinction of a signal traveling on the transmission medium. The attenuations are present on all mediums. The increase in attenuation can render the transmission inefficient or even inoperational.

In fiber optics, the increase of attenuation can cause malfunction in media converting equipments and failure to establish connectivity. Large fiber attenuation, which necessitates the use of amplification systems, can be caused by several factors such as:

- A combination of material absorption in the fiber optic cable either from poor technology or production faults, or because of physical stresses to the fiber or material ageing.
- Physical phenomenon like Rayleigh scattering that states that light or other electromagnetic waves traveling through transparent solid, liquid or gas materials can be scattered by particles that have a smaller size than the wavelength of the wave or like Mie scattering [1].
- Imperfect splicing either due to technology (mechanical splicing has higher attenuation) or due to poor quality splicing machine or mishandle.

The material absorption for silica, the major component in fiber glass, is around 0.03 dB/km. Modern fiber producing technologies only manage around 0,35 dB/km. The attenuation is also influenced by the wavelength of the light used as a carrier. Higher wavelengths experience lower attenuation: for example, at 1310 nm, in accordance to the ITU-G.652, the attenuation is 0,35 dB/km; at a higher wavelength of 1625 nm, the attenuation decreases to 0,25 dB/km.

Attenuation is also affected by the number of splices. In general telecom companies accept as a valid splice a splice that has an attenuation lower than 0,1 dB/km.

Copper wired and WiFi infrastructure modeling

a) Copper cables

Copper cables represent the standard method for interconnecting devices. Old cable types like narrow coaxial for token rings or other like technologies are not supported, instead focusing on the twisted pairs. The actual cable contains 4 pairs of twisted copper wires, hence the name: twisted pair. The wires inside are twisted to prevent crosstalk between neighboring pairs and to cancel out external electromagnetic interference.

There are two main classifiers of copper cables: category (CATx) and type (UTP/STP/FTP).

The cable's category is actually a design specification regarding its characteristics: impedance, propagation speed and delay, skew, maximum tensile load, wire size, cable thickness, bandwidth rating and so on. Mostly used categories are Cat5, Cat5e and Cat6. Cat5/Cat5e for example is meant for 100Mbps communication, while Cat6 is certified for 1Gbps.

The type of the cable is determined by its internal structure. UTP (Unshielded Twisted Pair) is the simplest cable type, containing only the four twisted pairs that are again twisted amongst themselves sheltered by a light plastic outer jacket. UTP is the most used cable type in the present due to its cheap cost and high flexibility, being used mostly indoors. FTP stands for Foiled Twisted Pair due to its extra metallic foil that covers the twisted pairs. This gives the cable a much stronger electromagnetic resistance, allowing safe usage both indoors and outdoors. The STP cable (Shielded Twisted Pair) offers even more protection, shielding each of the individual pairs with a conductive jacket. This shielding, while providing more EM and crosstalk proofing does increase its bulkiness and cost. STP is used mostly outdoors and in high EM environments.

The GIS takes into account these characteristics as the user needs to know where cables are installed and their type and category. For example, besides its physical position, it

is important to know if a failed link that will be replaced is a Cat6 STP or a simple Cat5e UTP cable. Moreover, Gigabit ports need to be interconnected with the correct cable type, otherwise the link will cause difficult to detect problems.

b) WiFi

Wireless technology plays an essential part of today Internet enabled networks. They allow users the possibility to move in a limited environment without losing network connectivity. Because accessing the wireless network can be done with relative ease this method of connecting the network has drawn many computer users especially the ones using laptops. This technology is used especially as a last mile solution. Although their benefits are notable there are several factors that make them unsafe for different security reasons and sometime unusable if not configured properly.

It is important to know the working frequency channels used by the WiFi enabled devices in order to ensure that they do not interfere. The proposed network management solution enables the network administrator to store information about the software configuration used for the equipment, monitor the activity of point to point links and equipments, bring up alerts when two access point devices operate on the same channel are too close together and thus might the effects of interference.

6. User interface

The user interface requirements are similar to all applications. The interface has to be clear and easy to be understood and used. It also has to be consistent throughout the application. The interface model is intuitive and resembles the way network technicians working in the field organize this type of data. Technologies like drag and drop are used to easily identify and link together transmission mediums and equipments.

7. Conclusions and future developments

The implemented network management system is able to accurately describe most of the components that make up the infrastructure of an internet data carrier and internet service provider. This is an extremely important issue for the Internet Service Providers that need to have up-to-date information about their infrastructure. Based on this information they can implement development strategies for updating and expanding the infrastructure with the help of the network management system. The importance of having a complete view of the infrastructure is also vital when troubleshooting network problems. In practice there is a big gap between the network administrators that are interested only in the logical configuration of the network, monitoring only network transmission performance parameters (round trip time, number of flows, packet loss, malformed packets such as runts, MTU, bandwidth) between hosts, ignoring altogether the physical layer, and the field technician that is responsible for maintaining the infrastructure from the physical point of view that often doesn't use any monitoring tools. The proposed network management system is designed to cover this gap by documenting the infrastructure. Based on that documentation and on the transmission parameter gathered automatically, an estimate can be given about how much of the physical infrastructure is working and how well. Using different algorithms

for spectral analysis as seen in [5] the information gathered from the network is synthesized and presented in an easily interpreted format.

In comparison to other solutions like the hardware ones using NoC (networks on chip) the technology presented in [6] is less expensive and offers a controllable degree of redundancy.

In accordance to the ISO / OSI network layers the proposed system covers the physical layer, data link layer and parts of the network layer. In the current configuration the system only analyzes the functionality at network layer assuming that the physical topology is the same with the logical topology. In real life in some cases the physical configuration can be totally different than the logical configuration. There are technologies such as: VLAN - virtual local area networks that allow logical level segregation, MPLS that is able to create "virtual links" between distant nodes, VPN - virtual private networks and IP tunneling technologies that allow two hosts to logically communicate as if they are alone in the network and directly connected. All these technologies do not affect the accuracy of the physical layer description but they affect the monitoring algorithm and also the prediction algorithm. Future improvements will be aimed at resolving these issues. Preprocessing of vector GIS maps with techniques similar to the ones presented in [7] can be added in order to reduce the complexity of the vector maps.

8. Bibliography

1. Anderson, R. D., Johnson, L. and Bell, F. G. **Troubleshooting optical-fiber networks: understanding and using your optical time-domain reflectometer**, 2nd edition, Amsterdam; Boston: Elsevier Academic Press, 2004
2. Antona, J.-C. and Bigo, S. **Physical design and performance estimation of heterogeneous optical transmission systems**, Comptes Rendus Physique, Volume 9, Issues 9-10, November-December 2008, pp. 963-984
3. Berry, J.K. **Beyond Mapping: Concepts, Algorithms and Issues in GIS**, Fort Collins, CO: GIS World Books, 1993
4. Ciordas, C., Hansson, A., Goossens K. and Basten, T. **A monitoring-aware network-on-chip design flow**, The Journal of Systems Architecture, 54, 2008, pp. 397-410
5. He, X., Papadopoulos, C., Heidemann, J., Mitra, U. and Riaz, U. **Remote detection of bottleneck links using spectral and statistical methods**, The International Journal of Computer and Telecommunications Networking, 53, 2009, pp. 279-298
6. Kolesnikova, A. and Fräntib, P. **Data reduction of large vector graphics**, Pattern Recognition, 38, 2005, pp. 381 - 394
7. Lee, H. and Baek, N. **AlexVG: An OpenVG implementation with SVG-Tiny support Computer Standards & Interfaces**, Computer Standards & Interfaces, Volume 31, Issue 4, June 2009, pp. 661-668
8. Smeureanu, A. I. and Dumitrescu, S. D. **Metode de imbunatatire a performantelor aplicatiilor GIS**, Conferinta anuala a doctoranzilor in stiinte economice, Academia de Studii Economice, 2009
9. Smith, M. J., Goodchild, M. F. and Longley P. A. **Geospatial analysis: A comprehensive guide to principles, techniques and software tools**, 2nd edition, Troubador Publishing Ltd., 2007
10. * * * http://en.wikipedia.org/wiki/Fiber-optic_communication#Attenuation

¹Acknowledgements

Some parts of this article are results of the project „Doctoral Program and PhD Students in the education research and innovation triangle“. This project is co funded by European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies (Project no. 7832, “Doctoral Program and PhD Students in the education research and innovation triangle, DOC-ECI”).

²Alexandru SMEUREANU graduated Politehnica University of Bucharest, Automatic Control and Computers Faculty, Computer Science Department and Bucharest University of Economics, the Faculty of Cybernetics, Statistics and Economic Informatics. He is currently a PhD candidate in the field of Economic Informatics at University of Economics. His interests range in the software programming, network management, GIS systems, vector graphics and embedded devices programming.

³Stefan Daniel DUMITRESCU graduated Politehnica University of Bucharest, Automatic Control and Computers Faculty, Computer Science Department. He is currently a PhD candidate in the field of Semantic Technologies at Politehnica University of Bucharest. His interests range in the informatics applications with accent on online development, information extraction, knowledge representation and human-computer interaction. Among other skills, he is also interested in project management and skilled in network technology.

⁴ Codification of references:

[1]	Anderson, R. D., Johnson, L. and Bell, F. G. Troubleshooting optical-fiber networks: understanding and using your optical time-domain reflectometer , 2 nd edition, Amsterdam; Boston: Elsevier Academic Press, 2004
[2]	Antona, J.-C. and Bigo, S. Physical design and performance estimation of heterogeneous optical transmission systems , Comptes Rendus Physique, Volume 9, Issues 9-10, November-December 2008, pp. 963-984
[3]	Berry, J.K. Beyond Mapping: Concepts, Algorithms and Issues in GIS , Fort Collins, CO: GIS World Books, 1993
[4]	Ciordas, C., Hansson, A., Goossens K. and Basten, T. A monitoring-aware network-on-chip design flow , The Journal of Systems Architecture, 54, 2008, pp. 397-410
[5]	He, X., Papadopoulos, C., Heidemann, J., Mitra, U. and Riaz, U. Remote detection of bottleneck links using spectral and statistical methods , The International Journal of Computer and Telecommunications Networking, 53, 2009, pp. 279-298
[6]	Kolesnikova, A. and Fräntib, P. Data reduction of large vector graphics , Pattern Recognition, 38, 2005, pp. 381 – 394
[7]	Lee, H. and Baek, N. AlexVG: An OpenVG implementation with SVG-Tiny support Computer Standards & Interfaces , Computer Standards & Interfaces, Volume 31, Issue 4, June 2009, pp. 661-668
[8]	Smeureanu, A. I. and Dumitrescu, S. D. Metode de imbunatatire a performantelor aplicatiilor GIS , Conferinta anuala a doctoranzilor in stiinte economice, Academia de Studii Economice, 2009
[9]	Smith, M. J., Goodchild, M. F. and Longley P. A. Geospatial analysis: A comprehensive guide to principles, techniques and software tools , 2nd edition, Troubador Publishing Ltd., 2007
[10]	*** http://en.wikipedia.org/wiki/Fiber-optic_communication#Attenuation

A MULTIFACTOR STATISTICAL MODEL FOR ANALYSING THE PHYSICO-CHEMICAL VARIABLES IN THE COASTAL AREA AT ST-LOUIS AND TAMARIN, MAURITIUS¹

Vandna JOWAHEER

Department of Mathematics, Faculty of Science,
University of Mauritius, Mauritius

E-mail: vandnaj@uom.ac.mu

Varunah LALBAHADOOR

Department of Mathematics, Faculty of Science,
University of Mauritius, Mauritius

E-mail:

Roshan RAMESSUR

Department of Chemistry, Faculty of Science,
University of Mauritius, Mauritius

E-mail:

Lutchmee DOSORUTH

Department of Chemistry, Faculty of Science,
University of Mauritius, Mauritius

E-mail:

Abstract: *The coastal pollution is an issue of concern for Mauritius. Since the past two decades, agricultural activities have contributed to pesticide and fertilizer run off in coastal waters. Over the recent years, the major urban and industrial growth in Mauritius have also contributed to water pollution by indirect wastewater discharge containing contaminants into rivers. As polluted water is hazardous to both marine life and human beings, it is of national interest to analyse the levels of water pollutants and the factors effecting these levels. This paper aims at developing a statistical model to evaluate the extent of coastal pollution in urbanized and agricultural regions in Mauritius. The study was carried out at two stations: St-Louis River (an urbanized industrial area) and Tamarin River (an agricultural area). A multifactor statistical model was formulated and analysed for the experimental data on the chemical and physical variables collected at the two sites for the estuary and downstream at each station, during 2001 and 2005 randomly spread over summer and winter seasons. The model is highly efficient in depicting the independent as well as the interactive effects of seasonality, time-interval, strategic locations and activities on the levels of different variables. A series of interesting conclusions were drawn from the analysis of the model. One major derivation was that the seasonal factor and time-interval had a significant effect (p -values < 0.01) on the levels of chromium, lead and nitrates at both the stations. However, the direction and magnitude were different with respect to each variable over the strategic locations. Moreover, considerable interactive effect between various factors regarding salinity was detected. These conclusions among others raise concern.*

Key words: *Experimentation; Statistical modeling and analysis; Estuaries; Trace metals; Nutrients; Physical parameters*

1. Introduction

Mauritius, an island geologically with a total land area of about 1865 square kilometres is dotted with rivers and streams where most of them are sourced in the high rainfall regions of the central plateau and flow down slope towards the Indian ocean. Due to the fact that Mauritius is so small, all land based activities have an impact on the coastal zone, thus making the coastal waters prone to pollution. Over the past years, the agricultural activities, the increase in industrialization and pressure on land due to high population density have produced large volumes of waste matter which have been indirectly dumped into the surrounding marine environments. The main sources of water pollution are contamination of underground water by pesticides and fertilizers, indirect disposal of industrial waste into streams and rivers, and contamination of domestic water supplies by overflowing of sewage system. (Ramessur et al., 2009; Burnett et al., 2006).

For over 40 years, about 45% of the existing land area was covered with sugarcane plantation. Over these years, a large amount of pesticides and fertilizers have been used in order to achieve a high yield but as side effects, these agricultural practices have represented particular risks to water sources. The fertilizers and pesticides have been penetrating the ground water sources and runoff during rainfall, thus adding to the level of contaminants in the surface waters. In addition, over the past two decades, several surface water bodies have been receiving industrial effluent discharges containing chemicals and trace of metals. Metal residues and textile slurries with high trace metal concentration have caused great deal of pollution. Hence there are strong reasons to believe that coastal waters of Mauritius may possibly be undergoing degradation of the water quality arising from the presence of biological and chemical pollutants. (Anon., 1998; Anon., 2005; Breward et al., 1998).

In order to study the physico-chemical activities in the coastal waters of Mauritius, several studies have been performed and the levels of various metals, nutrients, physical and chemical parameters were recorded (Ramessur et al., 1998; Ramessur et al., 2001; Ramessur 2004). These studies have shown the presence of the toxic metals such as chromium and lead. Since the pollution of rivers and estuaries is hazardous for both the marine life and human beings, it is important to find out whether the presence of toxic substances is significantly increasing due to the impact of industrial, agricultural activities, sewage and atmospheric pollution. The aim of this paper is to develop and analyse a multifactor statistical model to assess the extent of water pollution in the coastal areas of Mauritius where agricultural or industrial activities have been in practice over last few years.

2. Materials and methods

2.1. Study Sites

The 2 sites along the western coast of Mauritius extending from St Louis to Tamarin are shown in Figure 1.

St Louis Catchment urban estuaries (stations 1 and 2)

The Grand River Bay estuary receives urban runoff indirectly from the St.Louis River, which flows through Pailles and Plaine Lauzun industrial and urban area. The GRNW, which discharges south of Port Louis, has a catchment area of 116 km² and is fed by small southern tributaries from Upper Plaines Wilhems and Moka district and runs adjacent to the M1 motorway.

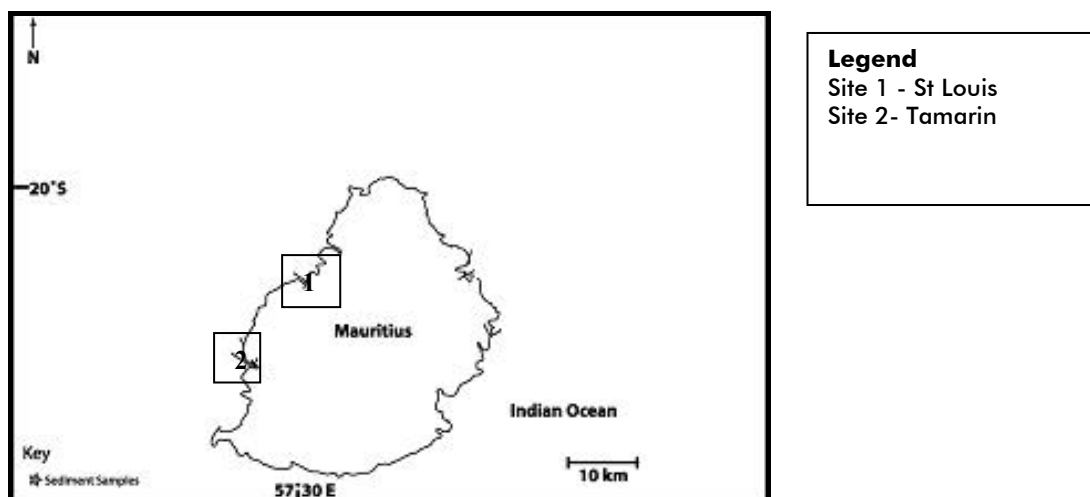


Figure 1. Sampling stations along 2 sites on the western coast of Mauritius

Note. Urban-site 1; Rural-site 2

Tamarin estuaries (stations 3 and 4)

The Tamarin estuary is unique in its kind where the topology of the sandy beach might change overnight. Sand deposits often block the mouth of the river such that water is trapped in the estuary forming a sort of basin. Anthropogenic activities that are near the estuary are mainly related to the tourism industry and also from human settlements upstream.

2.2. Experimental Data

Physical variables --- pH (v_1) & Salinity (v_2)

Salinity was recorded in the different compartments along the St Louis River using a Bellingham and Stanley field refractometer with a precision level of 0.2%. Replicate measurements of pH and temperature in situ the river were done using a Whatman pH portable meter A270. Replicate measurements of dissolved oxygen (D.O) were made in situ using a portable D.O meter Jenway Model 9071 consisting of a 'Clark' type polarographic oxygen electrode after calibration using a solution of 2% sodium sulphite.

Trace metal analysis in surface sediments ---- Pb (v_3) & Cr (v_4)

Cr and Pb were determined in the extracted solution from surface sediments collected along the 2 estuaries in 2001 and 2005 using a UNICAM 929 Atomic Absorption Spectrometer (AAS) (Analytical Technology Inc. 1993). Standards used for calibration for trace metal determination in sediments were prepared from standard 1000 mg L⁻¹ stock solution. The working solutions (10 mg L⁻¹) of each Pb and Cr were prepared by diluting the stock solution in a 100 mL volumetric flask with deionised water. Standards were then prepared in the range of 1.0-4.0 mg L⁻¹ for Pb and 1.0-5.0 mg L⁻¹ for Cr. Flame AAS used for trace metal analysis in surface sediments involved the use of a mixture of acetylene and nitrous oxide for Cr at a wavelength of 357.9 nm and an air-acetylene flame for Pb at a wavelength of 217.0 and 213.9 nm respectively. The sample of carrier gas flow rate was maintained between 200-500 mL min⁻¹.

Nutrients --- Nitrate (v_5), Nitrite (v_6) and Phosphate (v_7)

Replicate samples of water were collected in 200ml plastic bottles (dissolved nitrate) and glass bottles (reactive phosphate) in the coastal area in St Louis and Tamarin. Samples were stored at 4°C and analysed within 24h. The concentration of dissolved nitrite, dissolved nitrate, and dissolved phosphate were determined using standard spectrophotometric methods (Parsons et al., 1984) at 543nm and 882nm respectively using a PU 8710 spectrophotometer and a UNICAM 8700 Series UV/VIS spectrometer following calibration using known standard solutions. Quality control was achieved by analysing an internal reference independently prepared from the standard and the standard curves were verified after 10 successive runs by analysis of one standard solution within the linear range for each nutrient.

In the presence of mineral acids nitrites reacted with amines to give diazonium salts which in turn reacted with an organic amine to give a pink azo dye after 10min. The amount of azo dye produced was measured by its absorption of light at 543nm in a 10cm cuvette.

Dissolved nitrate in the samples were reduced almost quantitatively to nitrite by running samples and standards through a column containing commercially available cadmium granules coated with metallic copper. The nitrite produced were determined by diazotising with sulphanilamide and coupling with N-(1-naphthyl)-ethylenediamine dihydrochloride to form a highly coloured azo dye which was measured spectrophotometrically in 10 cm cuvettes at 543nm after 15 min. The reduction efficiency of the Cd column was determined by comparing the amount after reduction with the calculated amount supplied to the column. A correction was made for the nitrite present in the sample by analysing without the reduction step. The precision of nitrate determination was at the 1 mmol L⁻¹

Samples for reactive phosphate analysis were immediately filtered after collection. In a suitably acidified solution, phosphate reacted with molybdate to form molybdo-phosphoric acid, which was then reduced to the intensely coloured molybdenum blue complex after 5min. The absorbance of the latter was measured at 882nm in a 10cm cell. The precision was at the 0.1 mmol L⁻¹.

Quality Control

The accuracy and precision of the method was evaluated using three replicate determination of Standard Reference Material SRM 1646a Trace Elements in Estuarine Sediments from National Institute for Science and Technology (Colorado, U.S.A) which yielded 25.6 ± 1.8 mg kg⁻¹ (Cr); 8.7 ± 1.3 mg kg⁻¹ (Pb) compared to certified values of 40.9 ± 1.9 mg kg⁻¹ (Cr) and 11.7 ± 1.2 mg kg⁻¹ (Pb) respectively (mean and standard deviation) as determined using ICP MS by NIST with % recovery as 62.8%, 82.8% and 74.4% for Cr and Pb respectively. The main causes of losses were during the digestion procedure and also included random and systematic sources of uncertainty during analysis using the atomic absorption spectrometer. Necessary corrections were made accordingly. The detection limits were 2.3, 2.2 and 1.2 mg kg⁻¹ for Cr and Pb respectively. (The limits of detection were taken as 3σ of three replicates of the procedural blank digested filter paper (3x standard deviation about the mean).

2.3. Multifactor Statistical Model

We formulate a 2⁴ factorial model following Montgomery (1997), to investigate the effect of sites (A), years (B), stations (C) and seasons (D) each with two categories over each variable. The model equation is given by

$$y_{ijklr}^m = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ijl} + (\alpha\gamma\delta)_{ikl} + (\beta\gamma\delta)_{jkl} + (\alpha\beta\gamma\delta)_{ijkl} + \varepsilon_{ijklr} ; \quad i, j, k, l = 1, 2 ; r = 1, 2, \dots, n ;$$

where y_{ijklr}^m is the response from r^{th} sampling unit of m^{th} variable v_m ($m = 1, \dots, 7$) for the i^{th}, j^{th}, k^{th} and l^{th} categories of the factors A, B, C and D respectively; μ is the overall mean effect; $\alpha_i, \beta_j, \gamma_k$ and δ_l are the main effects; $(\alpha\beta), (\alpha\gamma), (\alpha\delta), (\beta\gamma), (\beta\delta)$ & $(\gamma\delta)$ are two-way interaction effects; $(\alpha\beta\gamma), (\alpha\beta\delta), (\alpha\gamma\delta)$ and $(\beta\gamma\delta)$ are three-way interaction effects; $(\alpha\beta\gamma\delta)$ is the effect of interaction between all the four factors and ε_{ijklr} is the error component corresponding to y_{ijklr} such that ε_{ijklr} are mutually independent and normally distributed with mean zero and variance σ^2 . The model fitting is done using the software SPSS 14.

3. Results and discussions

As depicted in Figure 2, exploratory analysis of experimental data indicates that the levels of lead and chromium are much higher in St-Louis as compared to Tamarin at both the sites and for both the time points. This can be attributed to the fact that since St-Louis river flows through Plaines Lauzan, an industrial zone, it may be receiving lots of industrial waste water which elevates the level of metal content in this river.

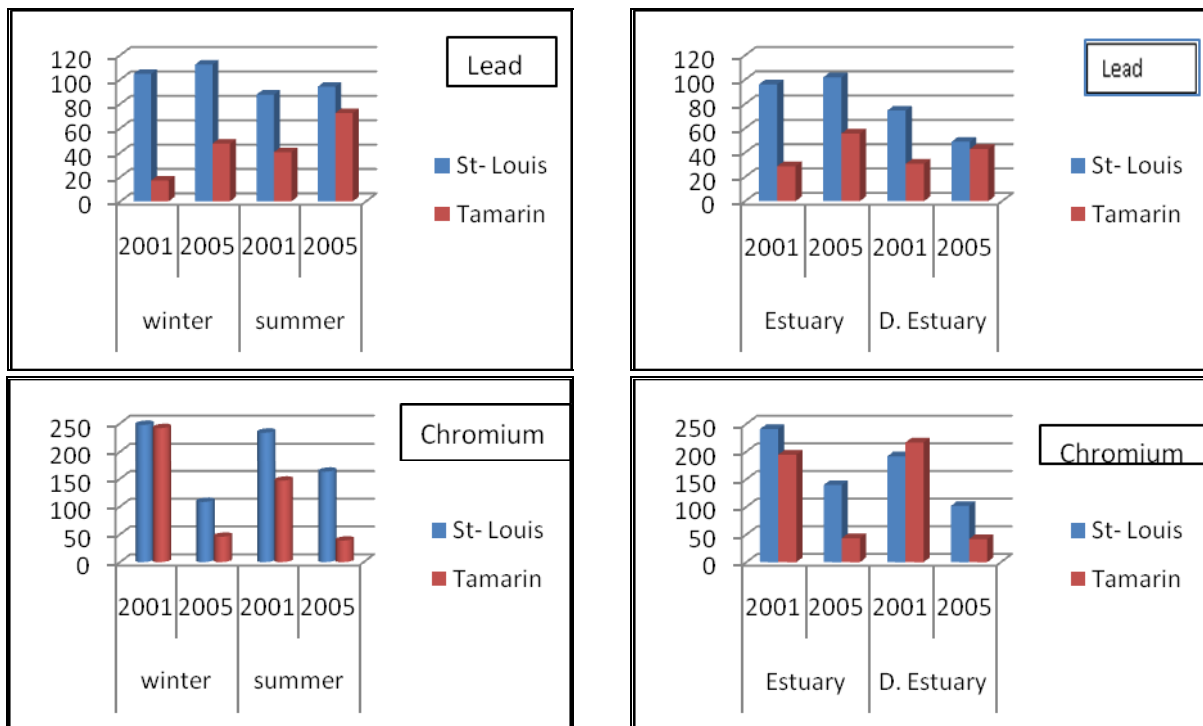


Figure 2. Distribution of Metal Contents

We also notice that in winter, the level of lead is higher at both the sites. From 2001 to 2005, there has been an increase in lead content and a decrease in chromium content at Tamarin. Moreover, estuaries have comparatively higher levels of lead than in downstreams.

Figure 3 reveals that salinity and pH levels are in general higher in Tamarin estuary and especially during the summer season, there appears to be a significant difference between the two sites concerning these two physical variables. However, in winter salinity is higher in St-Louis than in Tamarin over both the years. For both the sites, pH and salinity are much higher in downstream waters than in estuaries at both the time points. It is interesting to remark that pH levels have increased over years at both the stations of St-Louis whereas these levels show a decrease over years at both the stations of Tamarin.

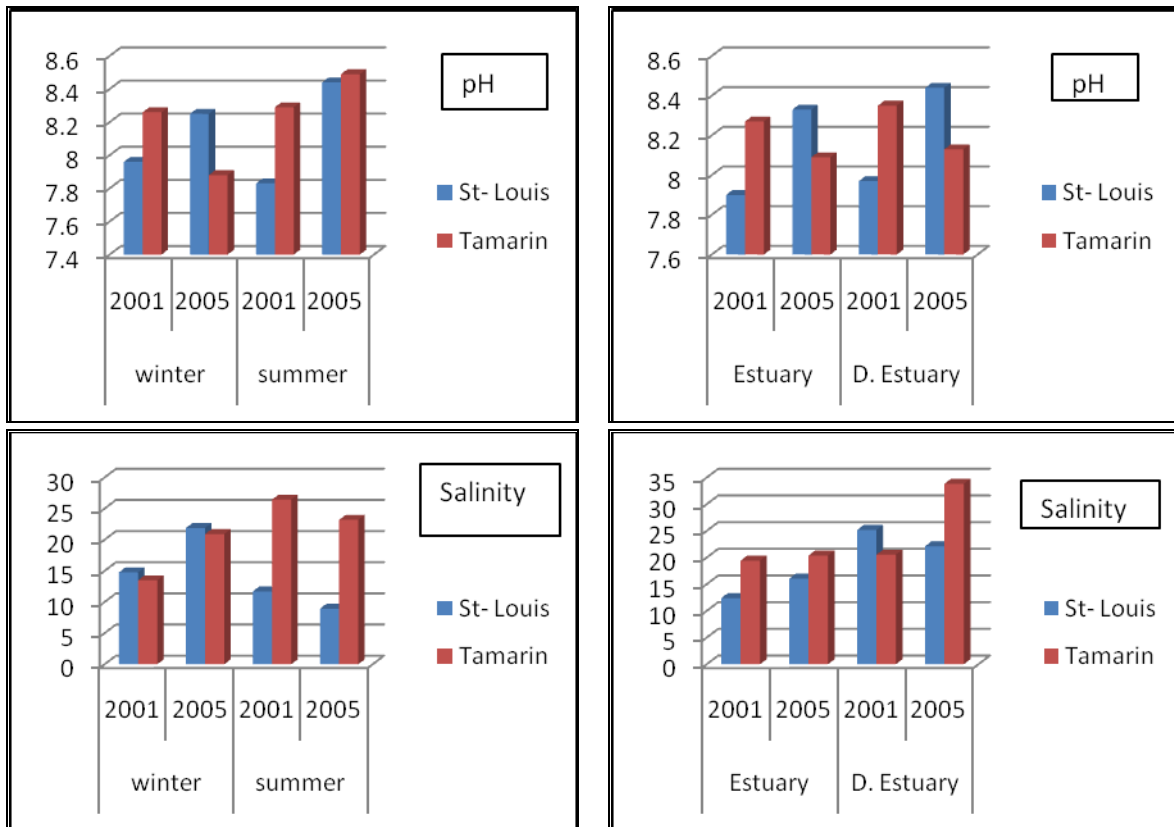


Figure 3. Distribution of Physical Variables

Concerning the level of nutrients, it is obvious from Figure 4, that the nitrate and nitrite contents at both sites have decreased drastically over the years. At St- Louis, this may be because industries are filtering their wastes before discharging into the river while at Tamarin, the decrease may be attributed to the fact that land being cleared for construction purposes, there is lesser cultivation, thus less use of fertilizers. It is observed that the phosphate content at St- Louis is higher except for winter 2005. Also, from 2001 to 2005 at Tamarin, we notice that there has been a general increase in the phosphate level.

Furthermore, the nitrate and nitrite content are higher at St- Louis at each station as compared to Tamarin over both years. Also, these nutrients at each site are higher at the estuaries. It is observed that the phosphate content at both stations was higher in St- Louis region in 2001 while at both stations in 2005, the phosphate content is higher at Tamarin. Overall, we notice that from 2001 to 2005, the phosphate content at St- Louis has fallen down while at Tamarin, there has been an increase. We also note that, the phosphate content at the estuary of each station is higher than that at downstream estuary.

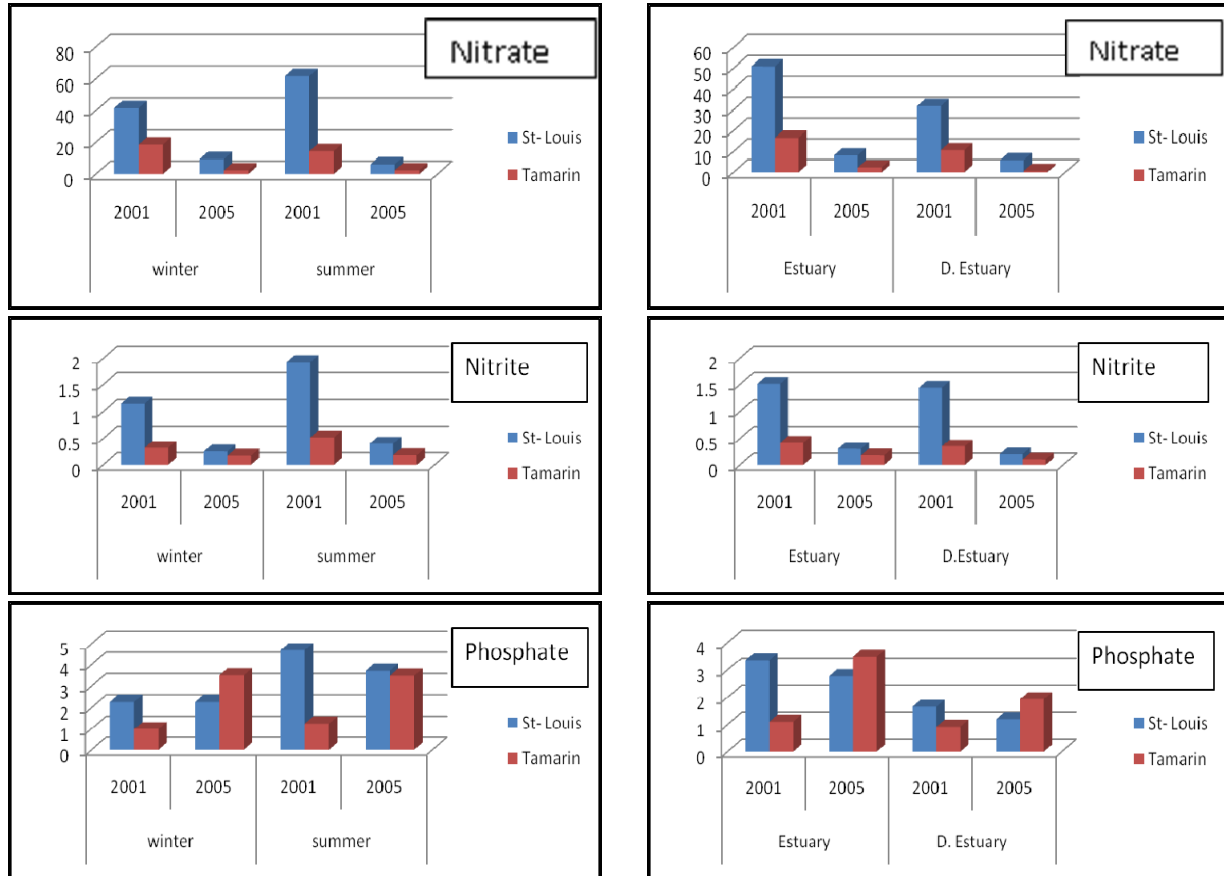


Figure 4. Distribution of Nutrients

With respect to the model fitting, all the assumptions of the model equation given in Section 2.3. are well satisfied by the data and the results obtained from model fitting are summarized in terms of p-values for the significant main factors as well as the interaction effects for all the variables under study and represented in Table 1.

Table 1. Significant p-values

Sources	Pb	Cr	pH	Salinity	NO ₃ ⁻	NO ₂ ⁻	PO ₄ ³⁻
A	0.000	0.001	-	0.000	0.000	0.000	-
B	-	0.000	0.003	0.036	0.000	0.000	-
C	0.009	-	-	0.000	0.016	-	0.014
D	-	-	-	0.016	-	-	-
AB	-	0.01	0.000	0.003	0.000	0.002	0.034
AC	-	0.034	-	-	-	-	-
AD	-	0.038	0.004	0.000	-	-	-
BD	-	0.005	0.001	0.000	-	-	-
ABC	-	-	-	0.000	-	-	-

There is a significant difference between the lead contents for sites St- Louis and Tamarin. Between the stations also, we observe a significant difference at 5%. The model analysis shows that St-Louis estuary exhibits greater levels of lead as compared to Tamarin estuary. Among the treatment combinations for chromium, we notice that there is a significant difference between the chromium content for the sites and for the different years. A significant interaction is found between sites and years, stations and sites, sites and seasons, years and seasons with respect to chromium.

Further, the level of chromium content for 2005 at St- Louis downstream estuary during summer significantly exceeds than that at Tamarin downstream estuary during summer.

We also note that there is a significant difference between the pH values for years. Moreover there is significant interaction between sites and years, sites and seasons, years and seasons, with respect to pH.

For salinity we can conclude that there is a significant difference between sites, years, stations and seasons. Moreover, there is significant interaction between sites and years, sites and seasons, year and seasons, and, stations, years and sites with respect to salinity. There is a significant increase in salinity during winter, both in estuaries as well as downstream estuaries and the levels are comparatively higher in St-Louis.

With respect to nitrate levels there is a significant difference between sites, years and stations. Also a significant interaction between sites and years is observed. The model analysis confirms that as compared to Tamarin, St-Louis has higher levels of nitrates at both the stations. However, when compared over years, the levels are lower in 2005 than in 2001. The same scenario is observed for the nitrate levels. Phosphate levels had been significantly higher in St-Louis river than at Tamarin in 2001 but the situation was reverse in 2005, which implies that phosphate levels have been considerably brought down at St-Louis over years.

4. Conclusions

The analysis demonstrate the potential for Pb concentrations in estuarine sediments at St Louis to exceed contamination limits. The contamination of sediments with Pb at St louis was significant showing an increasing trend over the years which could be considered to arise from accumulation and the heavy use of leaded petrol and potentially constitute a health hazard. However, storm runoff during flash floods in summer could cause a significant decrease in the levels of Cr in the rural estuary due to dilution with cleaner background sediments comparable. Excessive nutrients at St Louis and Tamarin can promote algal blooms, leading to oxygen depletion and severe deterioration of water quality as well as fish mortality. It can be argued that agricultural, urbanization and tourism activities have contributed to an increase in anthropogenic activities and hence have increased the potential for increased land and urban runoff in the two areas.

5. References

1. Anon **Mauritius Neap II: Environmental Strategy option report**, Environmental Resources Management London, 1998, pp. 173
2. Anon **Assessment and Management Implications of Submarine Groundwater Discharge into the Coastal Zone**, ICAM-IHP-IAEA, 2005, pp. 52
3. Breward, N. *et al.* **Manual for contaminant flux assessment in tropical coastal environments. Volume 1: Guidelines**, British Geological Survey, Overseas Geology Series, Technical report WC/98/41, 1998, pp. 56
4. Burnett, W.C. *et al.* **Quantifying Submarine Groundwater Discharge in the Coastal Zone via Multiple Methods**, Science of the Total Environment, Vol. 367, 2006, pp. 498–543
5. Montgomery, C. D. **Design and Analysis of Experiments**, 4th ed., John Wiley & Sons Inc, 1997, pp. 228-240
6. Parsons, T., Maita, Y. and Lalli, C. M. **Chemical and Biological methods for seawater analysis**, Pergamon Press, 1984

7. Ramessur, R. T. *et al.* **Statistical comparison between consecutive winter and summer concentrations in zinc and lead from sediments in a tropical urban estuary in Mauritius.** *Environmental Monitoring and Assessment*, DOI: 10.1007/s10661-009-1118-z, 2009
8. Ramessur, R. T., Parry, S. J. and Jarvis, K. E. **Characterization of some trace metals from the Export Processing Zone and a coastal tourist area in Mauritius,** *Environment International* 326, Vol. 24, No.7, 1998, pp. 773-781
9. Ramessur, R. T., Parry S. J. and Ramjeawon, T. **The relationship of dissolved Pb to some dissolved trace metals (Al, Cr, Mn and Zn) and to dissolved nitrate and phosphate in a freshwater aquatic system in Mauritius,** *Environment International*, Vol. 26 No. 4, 2001, pp. 223-230
10. Ramessur, R. T. and Ramjeawon, T. **Determination of Pb, Cr and Zn from an urbanized river in Mauritius,** *Environment International*, Vol. 28 No. 4, 2002, pp. 315-324

¹Acknowledgements

Thanks to Mr V. Ramsahye, Mr N. Ramsamy, Mr S. Mattapullut and Mr S. Radha for assistance during sampling work and analysis of sediment samples in the Chemistry Labs at the University of Mauritius.

THE USA SHADOW ECONOMY AND THE UNEMPLOYMENT RATE: GRANGER CAUSALITY RESULTS

Ion DOBRE

PhD, University Professor, Department of Economic Cybernetics
Vice- Dean of Faculty of Cybernetics, Statistics and Economic Informatics,
University of Economics, Bucharest, Romania

E-mail: dobrerio@ase.ro



Adriana AnaMaria ALEXANDRU

PhD Candidate, University Assistant, Department of Statistics and Econometrics
University of Economics, Bucharest, Romania

E-mail: adrianaalexandru@yahoo.com



Octavia LEPAS

Economist, HSBC, Paris, France

E-mail:

Abstract: *Using the time series data for USA shadow economy (SE), we examine the relationship between the size of unreported economy estimated as percentage of official GDP and the unemployment rate (UR). Granger causality tests are conducted, with a proper allowance for the non-stationarity of the data. The results indicate a clearly evidence of such causality from the unemployment rate to shadow economy.*

Key words: *shadow economy; unemployment rate; Granger causality*

Introduction

This paper uses the estimations of the U.S. shadow economy in order to evaluate if a structural relationship exists between the shadow economy and the unemployment rate for the United States. The structural relationship between the two variables is demonstrated by the use of an unrestricted VAR which shows the response of the shadow economy to a shock in the unemployment rate. The shadow economy is one of the causes of the inefficient functioning of the goods and labour markets. It introduces a distortion of competition within countries and among States. It is clear that the SE not only has negative effects on the economic system but also generates positive ones (Dell'Anno, 2007).

Shadow economy creates an extra added value that can be spent in the official economy. Schneider and Enste (2000) state that at least two thirds of the income earned in the SE is immediately spent in the official economy, thus having a positive effect on the latter.

The hidden economy expressed as percentage of measured GDP has been growing over the past of two or three decades. In many empirical and theoretical studies, it has been found that the tax burden is one of the biggest causes of the shadow economy, followed by the increase in government regulations such as through labour market regulations can lead to a huge increase in the cost of labour in the shadow economy.

Also, an increase in the unemployment rate reduces the proportion of workers employed in the formal sector. Consequently this leads to higher labour participation rates in the informal sector.

Data issues

In this context it is interesting to investigate the nature of the relationship between the unreported economy estimated as percentage of official GDP and the level of unemployment rate. The study use quarterly data for the period 1980-2007. Following the earlier work of Giles and Tedds (2000), Dell’Anno (2006) we have used the MIMIC models in order to generate quarterly data for the relative size of the USA shadow economy (Dobre and Alexandru, 2008).

We obtain the dimension of the shadow economy using an econometrical approach, in which we apply the MIMIC model with four causal variables and two indicators. Thus, we have obtained an MIMIC 4-1-2 as the best model, with four causal variables (tax on corporate income, social security contributions, unemployment rate, and self-employment) and two indicators (index of real GDP and civilian labour force participation rate).

For the unemployment rate the data are compiled from official data released by Bureau of Economic Analysis of USA (www.bea.gov).

We test each series for non-stationarity allowing for the possibilities of I (2), I (1) or I (0) data. To discover the unit roots, the Augmented Dickey-Fuller (ADF) test is used; to choose a number of lags sufficient to remove serial correlation in the residuals we have employed the Schwarz information criterion (ADF). In the following table the p-values is reported, while the null hypothesis is the presence of the unit root, and therefore a value greater than 0.05 indicates non-stationary time series.

Table1. Augmented Dickey-Fuller results

Variable		Level	First Difference	Second Difference
SE	None	0.1604	0.0006*	0.0000*
	C	0.7474	0.0056*	0.0000*
	T&C	0.1132	0.0290*	0.0000*
UR	None	0.2130	0.0000*	0.0000*
	C	0.2025	0.0000*	0.0000*
	T&C	0.2185	0.0000*	0.0000*

* means stationary at 0.05 level.

The results from table 1 indicate that both SE and UR are I(1), are hence non-stationary.

Is there a structural link between shadow economy and unemployment rate?

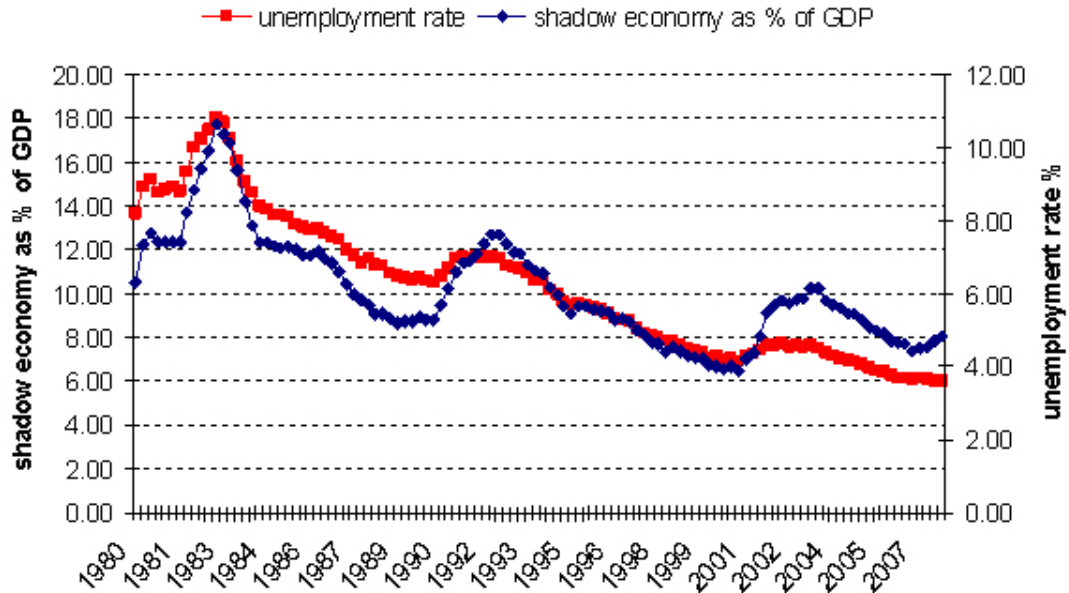


Figure 1. The shadow economy vs unemployment rate

The shadow economy measured as percentage of official GDP records the value of 13.6% in the first trimester of 1980 and follows an ascendant trend reaching the value of 18% in the last trimester of 1982. At the beginning of 1983, the dimension of USA shadow economy begins to decrease in intensity, recording the average value of 6.5% of GDP in 2007. The results of this estimation are not far from the last empirical studies for USA (Schneider and Enste 2001, Schneider 1998, 2000, 2004, 2007,). Schneider estimates in his last study¹, the size of shadow economy of USA as average 2004/05, at the level of 7.9 percentage of official GDP.

Figure 1 compares the trend of the shadow economy estimated by MIMIC model and the unemployment rate (UR) and shows a direct relationship between the two variables. The correlation between the estimated shadow economy and unemployment rate is found to be 0.90, confirming the presence of a strong positive relationship between the shadow economy and UR.

Giles and Tedds (2002) state that the effect of unemployment on the shadow economy is ambiguous. An increase in the number of unemployed increases the number of people who work in the black economy because they have more time. On the other hand, an increase in unemployment implies a decrease in the shadow economy. This is because the unemployment is negatively related to the growth of the official economy (Okun's law) and the shadow economy tends to rise with the growth of the official economy.

A general way of showing the relationship between the shadow economy and UR is to estimate an unrestricted VAR model. The optimal number of lags was chosen based on the Schwartz Bayesian Criterion and Akaike's Information. The optimal lag length was found to be 2, since the Schwartz, Akaike and Hannan Quinn information criterions indicates the same order of lag.

Table 2. The output of VAR model

	UR	SE
UR(-1)	1.346084	0.289852
	(0.15540)	(0.16672)
	[8.66220]	[1.73855]
UR(-2)	-0.427679	-0.378604
	(0.15810)	(0.16962)
	[-2.70510]	[-2.23205]
SE(-1)	0.200389	1.217077
	(0.14766)	(0.15842)
	[1.35713]	[7.68282]
SE(-2)	-0.173276	-0.190219
	(0.14926)	(0.16014)
	[-1.16087]	[-1.18783]
C	0.212894	0.201711
	(0.08841)	(0.09485)
	[2.40804]	[2.12660]
R-squared	0.980858	0.995224
Adj. R-squared	0.980129	0.995042
Sum sq. resids	4.555047	5.243034
S.E. equation	0.208282	0.223458
F-statistic	1345.097	5469.989
Log likelihood	19.05021	11.31366
Akaike AIC	-0.255458	-0.114794
Schwarz SC	-0.132709	0.007955
Mean dependent	6.082424	10.27581
S.D. dependent	1.477550	3.173557
Determinant Residual Covariance		0.000772
Log Likelihood (d.f. adjusted)		82.01566
Akaike Information Criteria		-1.309376
Schwarz Criteria		-1.063877

The estimated VAR is found to be stable (stationary), because all roots have modulus less than one and lie inside the unit circle. If the VAR is not stable, certain results (such as impulse response standard errors) are not valid.

Because the both variables are found to be integrated of first order, I(1) it is meaningful to test for possible cointegration between the two series and in table 3 we show the results applying the Johansen's likelihood ratio "trace test" to test the null of no cointegration in the context of a bivariate VAR model. Considering the inclusion of drift or trend in the VAR model, the five possibilities suggested by Johansen are considered. Asymptotic critical values are given by Osterwald-Lenum (1992).

Table 3. Johansen's "trace" likelihood ratio tests

	Drift/Trend Case ²				
	M1	M2	M3	M4	M5
	Trace Test Statistic(Ho:zero cointegrating vectors)				
Johansen's tests	19.53	27.22	9.40	15.82	15.69
Crit.value 5%	12.53	19.96	15.41	25.32	18.17
Crit.value 1%	16.31	24.6	20.04	30.45	23.46
	Trace Test Statistic(Ho:no more than one cointegrating vector)				
Johansen's tests	1.95	9.20	0.19	6.41	6.37
Crit.value 5%	3.84	9.24	3.76	12.25	3.74
Crit.value 1%	6.51	12.97	6.65	16.26	6.40

Assuming that we don't have deterministic trend in data, we see that we clearly reject the null of zero cointegrating, but cannot reject the null of one cointegrating vector.

In order to find the direction of any causality between UR and SE, we apply the Granger causality to the VAR estimated model. We estimate the system and then we can apply the usual Wald test to see if the coefficients of lagged SE variables are jointly zero in the UR equation. Similarly, we test if the coefficients of the lagged UR variables are jointly zero in the SE equation. In each case, the Wald test will be asymptotically Chi Square, with degrees of freedom equal to the number of "zero restrictions".

The results of applying the Wald tests for Granger non-causality to the VAR model appear in table 4.

Table 4: The Wald values of VAR Granger causality

Dependent variable: UR			
Exclude	Chi-sq	df	Prob.
SE	5.465254	2	0.0650
All	5.465254	2	0.0650

Dependent variable: SE			
Exclude	Chi-sq	df	Prob.
UR	9.793764	2	0.0075
All	9.793764	2	0.0075

The null hypothesis in each case is that the variable under consideration does not "Granger cause" the other variable. These results suggest that the direction of causality is from unemployment rate to shadow economy since the estimated Chi-square is significant at the 5 percent level; the critical Chi-sq=5.99(for 2 df). On the other hand, there is no "reverse causation" from shadow economy to unemployment rate, since the *Chi-sq* value is statistically insignificant.

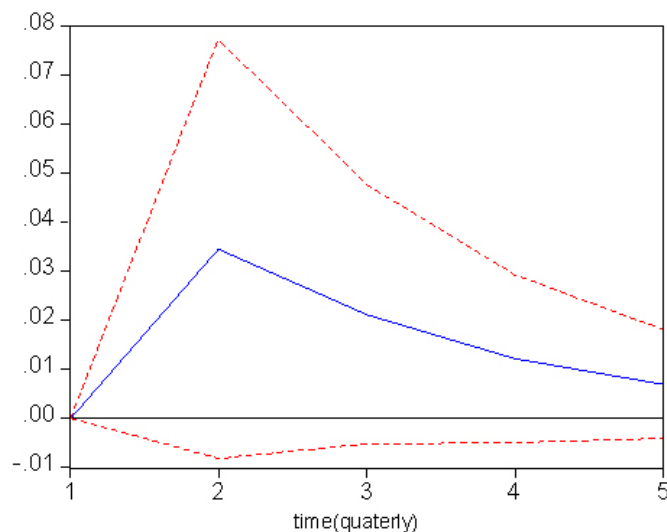


Figure 2. The response of shadow economy to a shock in the unemployment rate

Figure 2 shows that the shadow economy increases by about 3.5% above the baseline in response to a shock in UR. This is followed by a gradual decline towards the baseline. This occurs at the second quarter following the initial shock. This observation concurs with the theory that, an increase in the unemployment rate in the formal sector, fuels an increase in the number of people who work in the shadow economy. Consequently, there is an expansion in the size of the shadow economy.

Given that the estimation of the shadow economy, whose nature is unobservable, is very complicated, any theoretical and empirical inference derived by these figures should be considered always as an approximation.

Conclusions

In this paper we have used time-series data for the USA hidden economy and unemployment rate in order to explore the linkages between unemployment rate and the size of shadow economy in this country from 1980's to 2007. The size of the shadow economy was estimated using the 4-1-2 MIMIC model.

We find that the both series are cointegrated and there is a strong evidence of Granger causality from unemployment rate to shadow economy. On the other hand, there is no "reverse causation" from shadow economy to unemployment rate, since the *Chi-sq* value is statistically insignificant.

The impulse response function shows the response of shadow economy to a shock in the unemployment rate. Accordingly, shadow economy increases by about 3.5% above the baseline in response to a shock in UR. This is followed by a gradual decline towards the baseline. This occurs at the second quarter following the initial shock.

An increase in the unemployment rate in the formal sector, fuels an increase in the number of people who work in the shadow economy. Consequently, there is an expansion in the size of the shadow economy.

References

1. Aigner, D. J., Schneider, F. and Ghosh, D. **Me and my shadow: estimating the size of the U.S. hidden economy from time series data. Dynamic Econometric modelling**, "Proceedings of the Third International Symposium in Economic Theory and Econometrics", edited by Barnett, W., Berndt, E. and White, H., Cambridge University Press, 1986, pp. 297- 334
2. Dell'Anno, R. **Estimating the shadow economy in Italy: A structural equation approach**, Working Paper 2003-7, Department of Economics, University of Aarhus, 2003
3. Dell'Anno, R., Gomez, M. and Alañón Pardo, A. **Shadow economy in three different Mediterranean countries: France, Spain and Greece. A MIMIC approach**, Empirical Economics 33, 2007, pp. 51-84
4. Dell'Anno, R. and Schneider, F. **The Shadow Economy of Italy and other OECD Countries: What do we know?**, Mimeo, 2004
5. Dell'Anno, R. and Solomon, O. H. **Shadow Economy And Unemployment Rate In U.S.A. Is There A Structural Relationship? An Empirical Analysis**, The Annual Meeting of the European Public Choice Society, Finland, April 20-23, 2006

6. Dobre, I. and Alexandru, A. **Scenarii privind evolutia ratei somajului in Romania in perioada 2006-2009**, Revista Studii si Cercetari de Calcul Economic si Cibernetica Economica, 41/2 , 2007, pp. 39-52
7. Dobre, I. and Alexandru, A. **Posibilitati de estimare a dimensiunii economiei informale**, Revista Studii si Cercetari de Calcul Economic si Cibernetica Economica, 41/3, 2007, pp. 17-35
8. Dobre, I. and Alexandru, A. **Informal economy and USA unemployment rate. Is here is a structural relationship? An empirical analysis**, Macroeconomic of E.U. extension and economic growth Section, International Conference: Economic growth and E.U. Extension Process, Academy of Economic Studies, Bucharest, 2008
9. Dobre, I. and Alexandru, A. **The impact of unemployment rate on the dimension of shadow economy in Spain: A Structural Equation Approach**, paper presented at International Conference on Applied Business and Economics, Thessaloniki, Greece, 2008
10. Dickey, D. A. and Fuller, W. A. **Distribution of the estimators for autoregressive time series with a unit root**, Journal of Business and Economic Statistics, 15, 1981, pp. 455-461
11. Dickey, D. A. and Fuller, W. A. **Likelihood ratio statistics for time series with a unit root**, Econometrica, 49, 1981, pp. 1057-1072
12. Dolado, J. J., Jenkinson, T. and Sosvilla-Rivero, S. **Cointegrating and unit roots**, Journal of Economic Surveys, 4, 1990, pp. 249-273
13. Giles, D. E. A. **Causality between the measured and underground economies in New Zealand**, Applied Economics Letters, 4, 1997, pp. 63-67
14. Giles, D. E.A. **Measuring the size of the hidden economy and the tax gap in New Zealand: an econometric analysis**, Working Paper No. 5a, Working Paper on Monitoring the Health of the Tax System, Inland Revenue Department, Wellington, 1995
15. Gujarati, D. N. **Basic Econometrics**, 3rd Edition, McGraw-Hill, 1995
16. Johansen, S. **Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models**, Econometrica, 59, 1991, pp. 1551-1580
17. Johansen, S. **Likelihood-based Inference in Cointegrated Vector Autoregressive Models**, Oxford University Press, 1995
18. MacKinnon, J. G. **Numerical distribution functions for unit root and cointegration tests**, Journal of Applied Econometrics, 11, 1996, pp. 601-618
19. Schneider, F. **Shadow economies around the world: What do we really know**, European Journal of Political Economy 21, 2005, pp. 598-642
20. Schneider, F. and Enste, D. H. **Shadow economies: size, causes and consequences**, Journal of Economic Literature 38, 2000, pp. 77-114
21. Schneider, F. **Shadow Economies and Corruption all over the world: New estimates for 145 Countries**, Economics 2007/9, 2007, pp. 1-47
22. * * * www.bea.gov Bureau of Economic Analysis of USA
23. * * * www.bls.gov Bureau of Labour Statistics of USA
24. * * * Eviews Econometrics Computer Program:User's Reference Manual Version 5.1

¹ Schneider, F. **Shadow Economies and Corruption all over the world: New estimates for 145 Countries**, Economics, 9, 2007, 1-47

² M1-no drift/no trend in cointegrating equation or fitted VAR.
M2-drift/no trend in both cointegrating equation, no drift in fitted VAR.
M3-drift/no trend in both cointegrating equation and fitted VAR.
M4-drift and trend in cointegration equation, no trend in fitted VAR.
M5-drift and trend in cointegration equation and fitted VAR.

THE PHYSICS OF INFLATION: NEWTON'S LAW OF COOLING AND THE CONSUMER PRICE INDEX

Michael A. LEWIS^{1,2}

Hunter College School of Social Work, New York, USA



E-mail: michael.a.lewis@hunter.cuny.edu

Abstract: *In recent years, physicists have been using tools from physics to study social phenomena, an area of study sometimes called sociophysics and econophysics. Most of this work has appeared in physics journals, and the present paper is an attempt to bring this type of work to a largely social science audience. The focus is on the application of a differential equation model, widely used in physics, to the study of the long term trend in changes in the United States' price level. This model is found to provide an excellent fit to the data, indicating that this trend is an exponential growth trend.*

Key words: *inflation; price index; Newton's law of cooling; sociophysics; econophysics*

In recent years, physicists have been using tools from physics to study phenomena typically considered to fall within the domain of the social sciences, an endeavor sometimes referred to as sociophysics and/or econophysics. The work of physicists on social networks, collective decision making, financial issues, and income distribution (Toivonen, et al., 2006; de Silva, et al., 2006; and Clement and Gallegati, 2005; Galam, 1997) are some key examples. Although some social scientists have been open to this effort (Ormerod and Colbaugh, 2006 and Keen and Standish, 2006), I think it's fair to say that the influence of this work has probably been felt more within physics than outside of it. It is my view, however, that work on social scientific questions by physicists should be encouraged because the things we social scientists study are complex enough for us to need all the help we can get. I also think this approach of using tools from physics to study social issues should become more visible in venues that seem to mainly attract social scientists. My view about the importance of work in sociophysics and of moving it more out of the confines of physics venues is the occasion for this paper.

The paper focuses on the application of a differential equation model that's been used in financial theory and theories of economic growth but which comes up very frequently in physics and physical chemistry. Models of radioactive decay, Newton's Law of Cooling, and changes in the concentration of reactants over time for a first order reaction are three examples. What all these examples have in common is that the rate of change in some quantity over time is proportional to the amount of that quantity. When used to represent decay, the model stipulates that some quantity is decreasing over time. Thus, in radioactive decay the atoms of some element are decreasing in number over time, in first order

reactions the concentration of some reactant is decreasing over time, and in Newton's Law of Cooling the temperature of something is decreasing over time. However, Newton's Law of Cooling is also a "Law of Warming" when used to model increasing temperature over time. It's this last example of warming that's most relevant to the topic of this paper, since it's concerned with increase in the price level over time. In fact, inflation may best be conceived of as a kind of increase in the "temperature" of the macroeconomic system. In any case, it will be shown that the price level increases over time in a way analogous to increases in temperature in accordance with Newton's Law of Cooling. First, however, I will discuss previous work in economics on inflation.

Previous Work on Inflation

Much of the work in economics, that focuses on the dynamics of inflation, has been concerned with how changes in inflation are related to changes in other important macroeconomic variables such as unemployment, wage levels, and the money supply (Hess and Schweitzer, 2009; Christensen, 2001; Bjornrad and Nymoene, 2008;). There has also been work on the role of expectations and inflation. Rudd and Whelan (2005) provide a nice overview of this expectations oriented work. Those concerned about expectations and inflation have been mainly concerned about so called "rational expectations," a concept associated with New Classical Economists, and how this notion might be integrated into New Keynesian models of sticky prices.

Previous work in economics on inflation is important for both practical and theoretical reasons. Practically speaking, this type of work on inflation dynamics may lead to better forecasts of inflation that could be helpful in the design of macroeconomic policy. From a theoretical point of view, this work is important because it may move us closer to resolving theoretical debates about the factors that account for changes in inflation, debates among New Classics, Post-Keynesians, Keynesians, Monetarists, and others. The present paper, however, will "stand back" from previous work a bit and focus simply on mathematically modeling the overall long-term trend in the price level, deliberately neglecting many of the theoretical issues, that have preoccupied previous analysts. The only theoretical assumptions I make are 1) that increases in the price level lead to expectations of even higher increases in this level in the future and 2) that changes in the price level are proportional to changes in the level of aggregate demand. Based on these assumptions I propose a self fulfilling prophecy based theory of the long term trend in the price level.

If for some reason(s), perhaps including those discussed in the studies referred to above, the price level increases and if assumption 1 holds, this will lead potential buyers to expect an even higher price level rise in the future. This may result in the following outcome: aggregate demand will increase, as more buyers purchase more goods more quickly than was the case before the price level increase in an effort to act before the expected higher price level rise occurs. If assumption 2 holds, this will result in another increase in the price level. This increase will, in turn, lead to another, bigger than before, increase in aggregate demand, as, once again, more buyers purchase more goods more quickly that was the case before out of a desire to act before the next expected higher rise in the price level. What is being suggested here is what sociologists call a self fulfilling prophecy, a case where people's behavior, as a consequence of their expectations, end up causing the very thing

they expected to happen (Merton, 1968). In fact, as this process unfolds repeatedly, what develops is a kind of perpetual self fulfilling prophecy. In qualitative terms, such a process results in changes in the price level over time that are proportional to the level of prices at a given time. Below I translate this qualitative model into a mathematical one that's been inspired by a similar model in physics: Newton's law of Cooling.

Newton's Law of Cooling

A standard physics model attributed to Sir Isaac Newton, as a result of some experimental work he'd done, is known as Newton's Law of Cooling (Cengel, 2002). In equation form, the law states that:

$$dT/dt = k(T - T_s) \tag{1}$$

where "T" denotes the temperature of a given object, "t" denotes time, "T_s" denotes the temperature of the surrounding environment, and "k" is a constant of proportionality. Equation 1 is an example of an ordinary differential equation that can be solved by the method of separating variables (Stroud and Booth, 2005). If both sides of equation 1 are divided by (T - T_s) and multiplied by dt, we get:

$$dT / (T - T_s) = k dt \tag{2}$$

If we integrate both sides of equation 2, we get:

$$\ln(T - T_s) = kt + C \tag{3}$$

where "C" is an arbitrary constant. Taking the antilogarithms of both sides of equation 3 leaves us with:

$$T - T_s = e^{kt} + e^C \tag{4}$$

The rules of exponents stipulates that $e^{kt} + e^C = e^{kt}e^C$. Taking this into account and adding T_s to both sides of equation 4 gives us:

$$T = Ae^{kt} + T_s \tag{5}$$

Where $A = e^C$.

Equation 5 is the general solution of equation 1. If k is less than zero, equation 5 tells us how the temperature of an object, surrounded by an environment at T_s, will decrease over time until it reaches the same temperature as its environment. If k is greater than zero, equation 5 tells us how the object's temperature will increase over time, the more relevant case for the present paper. Notice in equation 5 that the increase or decrease in T is exponential to the point where the temperature of the object comes to equal that of the

surrounding environment. As will be seen below, this is the feature that's most analogous to the model of inflation this paper will focus on.

The mechanism behind Newton's law of cooling has to do with the second law of thermodynamics, which, in one formulation, stipulates that heat always flows from a higher temperature object to a lower temperature object (Arieh, Ben-Naim, 2008a and 2008b). Thus, considering equation 5, along with the second law of thermodynamics, k less than zero implies that, initially, the object is at a higher temperature than its surroundings. k greater than zero implies that the object is initially at a lower temperature than its surroundings. The microscopic interpretation of the second law of thermodynamics, a hallmark of statistical mechanics, explains the second law and, by extension, Newton's law of cooling in terms of the interactions of the particles that make up the systems under investigation (Arieh, Ben-Naim, 2008a and 2008b). The price inflation version of Newton's law, to be discussed below, also has a microscopic basis in the "particles" that make up the economic system. These particles are the buyers and sellers in the various markets that make up the economy, and I assume that these particles interact with one another in such a way as to lead to the kind of self fulfilling prophecy discussed above.

The Price Inflation Version of Newton's Law of Cooling

The type of feedback process involved in price levels changes, discussed above, has implications for how to model such changes. The self fulfilling process I've described leads to the stipulation that changes in the price level over time are proportional to the level of prices at a given time. Symbolically this is:

$$dP/dt = kP \tag{6}$$

Here "P" denotes the overall price level, "t" denotes time, and k is a constant of proportionality. This equation, like equation 1, can be solved by separating variables. If both sides of equation 1 are multiplied by dt and both sides are divided by P , we end up with:

$$dP/P = kdt \tag{7}$$

If we now integrate both sides of equation 7, we get:

$$\ln P = kt + C \tag{8}$$

where "C" is an arbitrary constant. Taking the antilogarithms of both sides of equation 8 leaves us with:

$$P = e^C e^{kt} \tag{9}$$

If we replace e^C with A we get:

$$P = Ae^{kt} \tag{10}$$

which is the general solution of equation 6.

Compare equation 10 with equation 5. There are very similar in form, with the only difference being that the right side of equation 5 has another constant, while the right side of equation 10 does not. However, both equations model exponential growth or decline, depending on the sign of k.

In order to see if equation 10, and by implication equation 6, fit available data, I took the logs of both sides of equation 10, which leaves us with:

$$\ln P = \ln A + kt \tag{11}$$

Data

In an effort to determine if equation 11 fit available data, I used data from the website of the United States Department of Labor (2009). This website contains monthly data on the Consumer Price Index for urban consumers (CPI-U) from 1913 to 2008 for a total of 1,164 data points. The CPI-U is calculated on a regular basis by U.S. analysts and is widely considered to be a measure of the general price level in the U.S. Thus, the CPI-U is my measure of P seen in equations 6 through 11; therefore instead of referring to the CPI-U below, for simplicity and consistency, I'll continue referring to P. Since I used monthly data, t in equations 6 through 11 is measured in units of months.

Results

Equation 11 was fit to the to the P series using ordinary least squares regression. The adjusted R² value for the model was .92, indicating an excellent fit to the data. That is, this R² value provides strong evidence that equations 6 through 10 more than adequately model inflation over this period, consistent with the self fulfilling nature of price level changes referred to above. The regression model output provided .003 as an estimate of k and 2 as an estimate of lnA. Thus, the price inflation version of Newton's Law of Cooling might best be described as a law of warming, with prices trending exponentially higher over time. It follows from the regression output that $A = e^{\ln A} = e^2 = 7$. Thus, equation 11 takes the form:

$$\ln P = 2 + .003t \tag{12}$$

and equation 10 takes the form:

$$P = 7e^{.003t} \tag{13}$$

Figure 1 displays a graph of P against t, indicating the exponential long term trend in P.

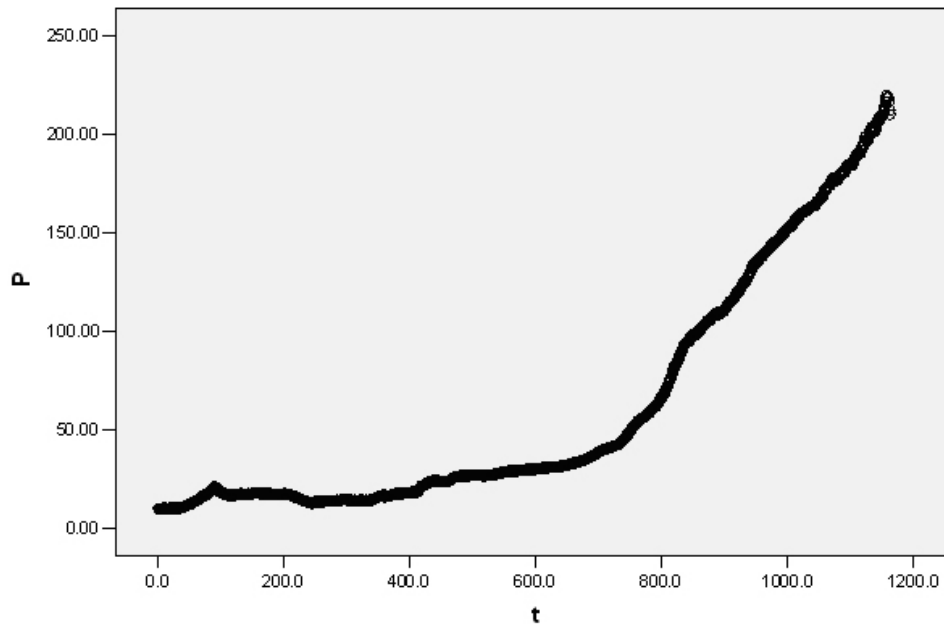


Figure 1. Graph of P against t

A key difference between this model and the physics version is that in the physics version, the laws of thermodynamics result in the temperature rise ceasing once equilibrium has been reached. There appears to be no laws of thermodynamics in the social realm to perform this function. Thus, aside from possible deflationary episodes caused by economic downturns, the upward trend in price level can apparently go on indefinitely.

The United States Department of Labor’s website also has price level data for the months of January through July of 2009. I used these data to test out of sample predictions of the model spelled out in equation 10. First, I calculated the predicted values for $T = 1,165$ through 1,171 (the months of January 2009 to July 2009) using equation 13. I compared these to the actual values for these time points from the Department of Labor’s website. The comparison was made using the percentage error formula which is:

$$\frac{(\text{actual value} - \text{predicted value})}{(\text{actual value})} * 100 \tag{14}$$

The R^2 value of .92 already provides evidence of the strong predictive power of the model, and this is reinforced by the fact that all of the absolute values of the percentage errors were less than 10%.

Discussion

This paper has focused on the analysis of changes in the price level. “Standing back” from traditional debates about the correlates of inflation, involving New Keynesians, New Classicals, Monetarists, and others, I have modeled the long term dynamics of price level changes on the assumption that increases in the price level respond to increases in aggregate demand and that price level changes unfold as a kind of self fulfilling prophecy. These assumptions led to an ordinary differential equation that’s popular in physics for

modeling the cooling or warming of objects in a surrounding medium, and this model was found to have an excellent fit to a time series of inflation data for the United States. Evidence for this fit was the very high R^2 value of .92 and the relatively small percentage errors referred to above. Thus, Newton's Law of Cooling appears to be applicable to the dynamics of price inflation and, hopefully, this finding will provide more impetus for and acceptance of the agenda of sociophysics.

References

1. ArieH, B.-N. **A Farewell to Entropy: Statistical Thermodynamics Based on Information**, New Jersey: World Scientific, 2008a
2. ArieH, B.-N. **Entropy Demystified: The Second Law Reduced to Plain Common Sense with Seven Simulated Games**, New Jersey: World Scientific, 2008b
3. Bjornstad, R. and Nymoen, R. **The New Keynesian Phillips Curve Tested on OECD Panel Data**, Using Econometrics for Assessing Economic Models, 2008, online source: <http://www.economics-ejournal.org/economics/journalarticles/2008-23>, 2009
4. Cengel, Y. A. **Introduction to Thermodynamics and Heat Transfer**, New York: McGraw-Hill, 2002
5. Christensen, M. **Real Supply Shocks and the Money Growth-Inflation Relationship**, Economic Letters, 22 (1), 2001, pp. 67-72
6. Clement, F. and Gallegati, M. **Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States**, Chatterjee, A., Yarlagadda, S. and Chakrabarti, B. K. (eds), *Econophysics of Wealth Distributions*, Milan: Springer-Verlag Itali, 2005, pp. 3-14
7. de Silva, R., Bazzon, A. L. C., Baraviera, A. T., Dahmen, S. R. **Emerging Collective Behavior and Local Properties of Financial Dynamics in a Public Investment Game**, Physica A Statistical and Theoretical Physics, 371 (2), 2006, pp. 610-626
8. Galam, S. **Rational Group Decision Making: A Random Field Ising Model at $T = 0$** , Physica A Statistical and Theoretical Physics, 238 (1-4), 1997, pp. 68-80
9. Hess, G. D. and Schweitzer, M. E. **Does Wage Inflation Cause Price Inflation?**, online source: <http://www.cleveland.org/research/PolicyDis/pd/.PDF>, 2009
10. Keen, S. and Standish, R. **Profit Maximization, Industry Structure, and Competition: A Critique of Neoclassical Theory**, Physica A Statistical and Theoretical Physics, 370, 2006, pp. 81-85
11. Merton, R. K. **Social Theory and Social Structure**, New York: Free Press, 1968
12. Ormerdod, P. and Colbaugh, R. **Cascades of Failure and Extinction in Evolving Complex Systems**, Journal of Artificial Societies and Social Simulation, 9 (4), 2006, online source: <http://jasss/soc.surrey.ac.uk/9/4/9/9.pdf>
13. Rudd, J. B. and Whelan, K. **Modeling Inflation Dynamics: A Critical Review of Recent Research**, 2005, online source: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=874757, 2009
14. Stroud, K. A. and Booth, D. **Differential Equations**, New York: Industrial Press, Inc., 2005
15. Toivonen, R., Onnela, J. P., Saramaki, J., Hyvonen, J. and Kaski, K. **A Model for Social Networks**, Physica A Statistical and Theoretical Physics, 371 (2), 2006, pp. 851-860
16. The United States Department of Labor, online source: <ftp://ftp.bls.gov/pub/special.requests/cpi/cpi.txt>, 2009

¹ **Correspondence:** Michael A. Lewis, 129 East 79th Street, New York, NY 10075

²Michael Lewis holds a masters degree in social work From Columbia University and a doctorate in Sociology from the City University of New York Graduate Center. He teaches courses in social welfare policy and political economy at the Hunter College School of Social Work. Lewis' research interests are in poverty, social welfare policy, and quantitative methods. His work, some co-authored with Eri Noguchi, has appeared in a number of peer-review journals.

METHODS OF PORTFOLIO MANAGEMENT FOR LISTED SHARES. SOME FEATURES FOR THE ROMANIAN PRIVATE PENSION FUNDS¹

Mihaela DRAGOTA

PhD, University Professor, Department of Finance,
University of Economics, Bucharest, Romania

E-mail: mihaela.dragota@fin.ase.ro

Natalia SUSANU

Master Student, DAFI,
University of Economics, Bucharest, Romania

E-mail: natalia.susanu@gmail.com

Abstract: *Recently, Romania put into practice the private pension system, which includes compulsory pensions and voluntary ones, as alternatives to statutory (public) pension scheme. According to the Romanian laws, private pension funds can invest in securities issued on regulated and supervised market from Romania, EU Member States, European Economic Space, and third countries a percentage ranging between 10% and 50% of the total pension fund assets. This study should be viewed in the current global financial context whereas, currently, the most powerful and stable financial markets are experiencing problems with the sudden drop in financial asset prices, low liquidity and a reduction in investment on the capital market. The administrators of these funds for the fundamental analysis applied on securities should consider the social component of their activity, since the public pension system is undercapitalized and encountered many problems arising from the influence of economic, social, demographic, political factors.*

In this article, for the selection of listed stocks to be included in a portfolio, we propose a score function. Using this method we determine which of the companies analyzed previously had the best performance in terms of an investor with risk aversion, and the final goal will be identify the best three companies included in BET index in terms of return and risk. The weights of each indicator were the results of the use of Likert scale.

Key words: *private pension system; portfolio management; investments; score function; Likert scale*

1. Private pension system in the European Union. The case of Romania

In the year 2008, Romania put into practice the private pension system, which includes compulsory pensions and voluntary ones, as alternatives to statutory (public) pension scheme, regarded as inadequate for social protection of the active populations, became further pensioners. The public pension system had significant problems, not only in Romania, but in the entire European Union. This is the reason why all the states tried to find alternative solutions,

having both similar strategies and some national specific measures, based on some qualitative and quantitative indicators which ranked the countries from this point of view.

Private pension funds in Romania establish their activities on a well regulated system and their performances will depend on the future private compulsory pensions. Taking into consideration the social component of these types of funds for the active employed population upon 35 years, the investments of insured population contributions are strictly regulated by the Romanian laws.

Private pension system in Romania was built in accordance to World Bank classification for this field, named **pillar of pensions**²:

- Pillar I – public pension systems, as type “pay as you go”, PAYG, with *defined benefits*;
- Pillar II – privately administrated pension funds, with *defined contribution*, compulsory for young population with age under 35 years and voluntary for active persons with age between 35 and 45 years;
- Pillar III – voluntary pension systems, private administrated, based on individual accounts.

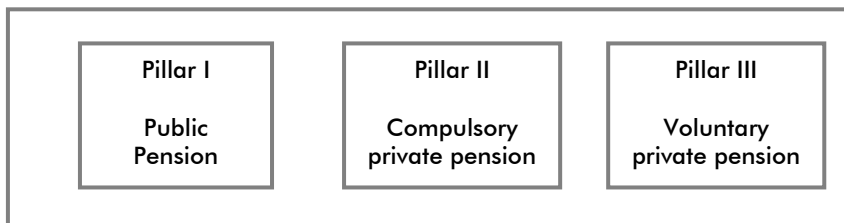


Figure 1. The multi-pillar pension system built and implemented in Romania since 2008

In the European Union, were identified four categories of states, for the implemented pension system point of view, respectively:

1. with less developed private pension systems, with no intentions to change these circumstances: Spain and France;
2. with well developed private pension systems, based constantly on these systems: Denmark, Holland, Great Britain;
3. with public and reformed “pay as you go” systems, with compulsory private pillar pension, supporting the public system with significant difficulties: Bulgaria, Estonia, Latvia, Lithuania, Hungary, Poland, Romania, Slovakia, Sweden;
4. with traditional social insurance systems, sometimes together with a minimum level of social insurance, this had implemented a private pension system: Germany, Austria, and Italy.

The multipillar system can be described having some particularities for Central and Eastern Europe countries, compared to more-developed states from European Union, as it can be seen in the following table:

Table 1. The multipillar pension systems from the European Union. Particularities for Central and Eastern Europe

	PILLAR I	PILLAR II	PILLAR III	PILLAR IV
Central and East Europe	Public pension	Compulsory private pension	Voluntary private pension	Occupational pension
UE 15 countries	Public pension	Occupational pension	private pension	

Source: www.csspp.ro

The starting point for multipillar pension systems takes into consideration the demographical characteristics of Europe, a continent which slowly, but surely, gets old. The decrease of live birth per year and the fertility rate, together with improvements in life expectancies in some countries are the most important reasons. Each member state must build an individual pension system, taking into account different previous crises – political and economical – sometimes with armed conflicts, having distinct developing level of the economy.

For instance, Croatia is one of the countries with an eventful recent history, with significant implications on the national pension system. In almost one decade (1991-2001), its population decreased with 3% (in absolute values, approximately 150.000 inhabitants), in the year 2001, getting to 4.300.000 citizens. The demographic structure was changed due to the armed conflict in this geographic region. The number of pensioners was growing, including war invalids and the descendants of the victim of the conflagrations (for example, war widow). Moreover, the deceased persons in this conflict are social contributors for the pay as you go system, which were diminished with more than 30% (for 525.000 to 360.000 persons). This special case sustains the requirement for methods based on socio-economic features for each country³.

In Romania, the most important reasons for pension reform are: the population ageing, the increasing values of old age dependency ratio, with possible insolvency of the “pay as you go” pension system. In order to improve the current public system, including the case of Romania, were taking into consideration some direct measures such as: increasing of the age for retirement, but together with special measures for those who require early retirement, because the population can choose, for example, advanced or invalidity pension and can work in the same time, based on the cost of opportunity concept. Regarding the age of retirement, there are a lot of public debates in the European Union referring to the gender differences for normal-age public pension, even if the life expectancy are relatively similar for both men and women. In some countries from European Union, the age for retirement has already been equalised or will be progressive equalised in the future (for example, Estonia will equalised the age for retirement until 2016, at 63 years).

In Romania, the multipillar system can be seen as a possible solution to prevent the insolvency of the public pension system. Especially, the pillar III can be an alternative solution, but, for example, in 2008 this sector is still not developed. The annual yield for private pension funds were between 6% and 50.83%. The most important causes for this evolution are the legal provisions, such as the tax deduction of social contribution for voluntary private pension: 250 euro, in real terms means 40 euros (15%·250 euros). In Czech Republic, the level of tax deduction is around 12.000 Czech crowns, which it means about 480 euros (20 November 2008⁴). Moreover, can be mentioned the limit of 15% from gross income for saving placements in pension pillar III.

2. Investment policy for private administrated pension fund in Romania

Private pension system in Romania was elaborated as defined contribution system, which it means that the insured person knows the percent/the level of paid contribution for his entire active life, without knowing the level of benefits as private pension, after the retirement. The opposite system has the definite benefits, the social contributor knowing the level of his future pension, based on his regular payments as social contribution.

According to the Romanian laws, both Pillar II and III guarantee, as a private pension, a sum equal to total paid contribution, less transfer penalties and legal services⁵. In order to multiply the total contribution paid by the social contributors, the pension fund administrator must invest the assets, taking into consideration the social role of this type of institutional investor.

The investments made by an administrator, taking into consideration a medium level for risk, with a less impact of profitability, must comply with some legal limits. The most important financial instruments (with maximum limits of 70% and the minimum percent of 10%) in which a private administrated pension fund can invest are as it follows:

- deposits in rol or hard currency to any credit institution authorized to function in Romania, in European Union or in European Economic Space – maximum 20% in total assets;
- accounts in rol or hard currency to any credit institution authorized to function in Romania, in European Union or in European Economic Space – maximum 5% in total assets;
- T-bonds issued in Romania or in any European Member state or in European Economic Space – maximum 70% in total assets, with the following sub-limits: (1) 50% in T-Bonds with a maturity less than a year; (2) 70% in T-Bonds with a maturity more than a year;
- Bonds and other financial assets issued by the local authorities from Romania or any European Member state or in European Economic Space – maximum 30% in total assets;
- securities traded on regulated and supervised capital markets from Romania, EU Member States or the European Economic Space - maximum 50% in total assets with the following sub-limits;
 - securities traded on regulated markets in Romania - 35%;
 - securities traded on regulated markets from the European Union or European Economic Space, other than Romania - 35%;
 - corporate bonds of the Romanian issuers - 30%;
 - corporate bonds from issuers from the European Union Member States or from the European Economic Area other than Romania, which received from international rating agencies the rating "investment grade" - 30%.

Initially was not taken into account the information concerning the market where the securities are traded - in Romania, any other Member States from the European Union or European Economic Space and the percentage of 50% was considered the maximum for all the assets required. The Regulation no. 3/2009 on the investments from the private administrated pension funds the interpretation was diversified and the emphasis is mainly on ratings given by international rating agencies. We can observe that the regulation did not say anything about the corporate bonds with no rating as "investment grade".

- T-bonds and other financial instruments issued by third countries – maximum 15% from the total assets;
- bonds and other securities issued by local government authorities in third countries, the percentage of up to 10% of the total pension fund assets etc.

In the period 2008-2009, we can not made statistically significant considerations about the yields of the privately managed pension funds from Romania because it would be simple speculation. It can make some observations when we analyze the portfolio-target for Pillar II and III, but it is possible that the deviation is too large, as a factor influencing this observation being also the short time horizon, for which we have no information.

According to official data from Private Pension System Supervisory Commission (PPSSC), at the end of the August 2009, especially due to the financial and economic crisis, the most

important destination for investments were government securities (58.37%), about 35% from the total assets were invested in corporate bonds, municipal bonds, supranational bonds and shares and 4.66% in bank deposits⁶.

Regarding voluntary pension from Pillar III, at this moment their portfolios are targeted mainly to government bonds, approx. 60% from their total assets. At the end of August 2009, from the total amount of assets of 161.5 million rol, administrated by the voluntary pension funds, 14.28% were placed in corporate bonds, 11.07% in shares, 5.06% in municipal bonds and 3.61% in bank deposits.

Here are two examples of states that have implemented a few years the multipilon pension system and they already have publicly available information about the return of private administrated funds. Thus, at the end of 2008, Czech Republic, with a population of 10,241,000 inhabitants and a GDP of 198.978 billion USD have one of the most powerful markets of private pension funds. Net assets from the Pillar III recorded a remarkable upward trend, coming from a volume of 6.342 million crowns, made in 1995 to 167.196 million crowns in 2007, i.e. an increase in nominal terms by 2636%. Over 40% of the Czech Republic has a private pension account, at June 30, 2008 there are 4,135,169 of participants and the market leader, as in Romania, is the ING company, with a market share of 27.03%.

Hungary is also a reference market for private pension funds. For the third pillar, the number of adherents increased continuously for the period 1998-2007, coming in the last year of that period to 1,385,440 members, a very large number if we think that their population is about 9,981,330 inhabitants. Total assets ranged from 100.41 million forints (in 1998) and 744.37 million forints (value recorded for 2007)⁷.

In February 2009, PPSSC came with new regulatory provisions (Regulation no. 3/2009) regarding **the degree of risk associated with privately administrated pension fund**. Depending on the proportion of the low-risk instruments in their portfolios, each of the Romanian private administrated pension funds may be associated with one of the following degrees of risk:

- low risk – with a total holdings of the low-risk instruments varying between 100% -85%;
- medium risk - with a total holdings of the low-risk instruments varying between 85% - 65%;
- high risk - with total holdings of the low-risk instruments varying between 65% - 50%.

In conclusion, we can say that despite some legislative provisions that provide some guarantees regarding the granting of private pensions when the participants of today will meet the retirement age, the uncertainty regarding the level of these pensions is obvious. Supervisory Commission of Private Pension System has a great responsibility from this point of view, the specific legislation being emerging.

3. Methods of portfolio management based on listed shares on Bucharest Stock Exchange

As shown in the previous subsection, private pension funds can invest a percentage ranging **between 10% and 50% of the total pension fund assets in securities** issued on regulated and supervised market from Romania, EU Member States, European Economic Space, and third countries. The percent of 50% of the assets is quite high, which may lead a private pension fund to pass from a medium risk level to one higher. In Romania we can illustrate with

such a pension fund that is Generali, which announce the target structure of its portfolio is considering investing 36% of its assets in securities traded.

If we analyse the problem of setting up a portfolio which can contain stocks, in order to assure a certain percentage of growth without excessive increase risk in the global financial crisis, we can say that the portfolio manager of any pension fund is facing a major challenge at least in the short to medium term. For this reason, we consider that this research approach is particularly useful as the capital markets worldwide are in the process of reconfiguring and resizing, being strongly marked by decreases in exchange rates, even for the best of financial investment.

We focused on Romania as a country that is emerging and the Romanian capital market is a market with a significant potential of growth. This fact is demonstrated by the data provided by a report from the agency Standard Poor's. According to this report, during 2002-2006, the Romanian capital market recorded an annualized increase of 69.7% while the S&P/IFCG index has in the same period an increase of 36.2%. The Romanian Stock Exchange recorded higher returns over other emerging markets from Russia, China, Egypt and Brazil. After the EU accession, the Romanian capital market recorded a growth of 15-25% in 2007.

This study should be viewed in the current global financial context whereas, currently, the most powerful and stable financial markets are experiencing problems with the sudden drop in financial asset prices, low liquidity and a reduction in investment on the capital market. This is confirmed by high correlation between the BET index and S& P500 of 0.92, recorded last year. Therefore, the distrust of investors in capital markets is generalized in an international context.

4. Some recent financial trends of the Romanian capital market

In Romania, the first steps regarding capital markets' creation were taken upon the establishment of trade exchanges in 1839 and Bucharest Stock Exchange (BSE) was officially opened on December 1st 1882. Its activity was influenced by social and political events of the time (the uprising of 1907, 1912-1913 Balkan War, World War, when the stock market was closed), and in 1948 the Effects, Shares and Exchange Stock Exchange closed down. Stock Exchange has started again, when BSE was re-established⁸.

The evolution of the listed shares was moderated in the early trading according to the consolidation and implementation of the mechanisms for trade and investment and due to the lack of investor's confidence in the Romanian capital market, and after that it had an upward trend from 1996 to 2007. Thus, in 1996 there was a decrease in market capitalization, with an upward trend in 1997 through the adoption of program trading for 5 days per week, increasing the number of issuers and financial intermediaries, but more importantly, due to massive foreign investment on the Romanian capital market. In 1998 followed the market capitalization dropped again due to lack of investor confidence determined by the political instability. Since 1999, this indicator has seen an upward trend, moving from 572.5 million rol in 1999 to around 85 billion lei in 2007, registering an increase of about 145.64%.

The year 2007 is the last year in which BSE has increased (see figure no.2), and after that followed a significant decrease for the most important stock indexes and indicators. Thus, BET reported in 2008 the largest annual decrease recorded in its entire existence, approximately 70% decrease, which cancelled the increases from the previous years. Also, in February 2009, BET has the lowest value from October 2003 to date, of 1887.14 points. Market capitalization in 2008 had a reduction of approximately 46% from the previous year and the total turnover

decreased by approximately 49% over the same period. Surprising, in September 2009 compared with the previous year, the market capitalization has seen an increase of 76%, but the total turnover decreased with almost 45%. These trends are explained due to the increase of sales from the part of the non-resident investors, but in the last two years their purchases of shares on the Romanian capital market fell. Instead, residents have performed with speculative investments, leading to increased volatility on the exchange market in Romania. The evolution of market capitalization and total turnover are similar to that seen on the emerging markets in the area.

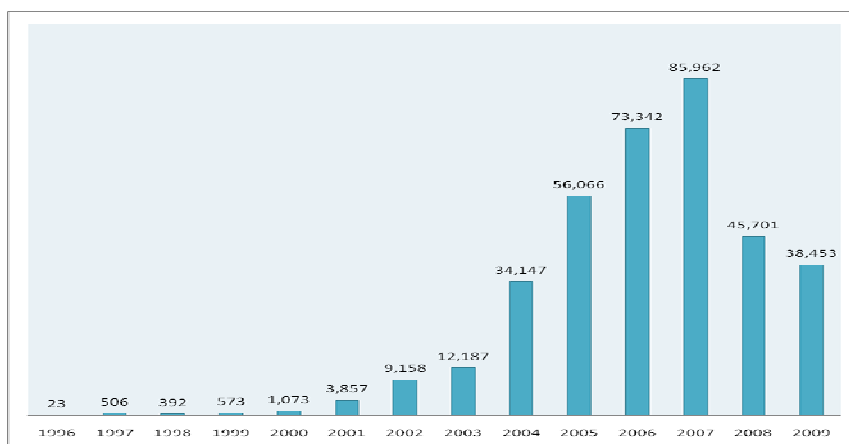


Figure 2. The evolution of the market capitalization on BSE during the period 1996-2009 (mil. Rol)

Source: <http://www.bvb.ro>

5. Methods of fundamental analysis applied on shares issued by the Romanian listed companies

Fundamental analysis applied on securities has two main components, namely an analysis using both financial, accounting indicators and capital market ones. The two types of analysis are based on specific indicators, but it is recommended to use both methods to obtain a more accurate economic and financial reality for each issuing company.

Private pension funds must make a very careful analysis, both regular and more complex of the portfolios of financial instruments created in order to make the necessary corrections in time when economic reality requires. This type of analysis should consider the social component of the activity of each private pension funds from Romania (and beyond) since they were created as a viable and more profitable alternative than public pension system, undercapitalized and encountered many problems arising from the influence of economic, social, demographic, political factors.

While private pension funds can invest in Romania up 50% of assets in shares, we realized the next study in terms of investor with risk aversion due to the abovementioned arguments. Were used financial information provided by the selected companies for the period 2004-2007, and stock indicators registered by them in 2005-2009. To take into account the important feature of a stock portfolio of a private pension fund, we choose the assessed shares to be those included in the structure of BET index. This selection is relevant because BET index includes ten most liquid shares traded on BSE. In a first phase, will examine the economic and financial situation for the selected companies and, in a second phase, based on the results

obtained from the scores' method, will be chosen the top three companies whose shares can be recommended to an investor with risk aversion.

5.1. Analysis of financial ratios for shares add in the BET Index using the fundamental analysis

The most important five financial ratios based on the Balance Sheets of the ten companies included in the BET Index are represented in the Annex 1. There is also included the inflation rate which helps us to evaluate the indicators in real terms.

The figures from Annex 1 show us that the majority of these companies have reported values for return on assets and return on equity lower than the inflation rate in the year 2004. Still, in the next years, the situation has improved and these ratios have overcome the inflation rate. Besides, a common feature of these companies is represented by the higher level of return on equity compared with the level of return on assets, which highlights the fact that debts facilitated to gain due to the leverage effect.

Investors with risk aversion are more likely to give a higher importance to indicators like current liquidity, solvability rate and total debt/total assets because these indicators reflect the companies' ability to cover the current debt, to pay the obligations on long term and they also reflect the proportion of the external financial resources in the stable financial resources⁹.

In this context, our study is relying on the evolution of these ratios, as it follows:

- *Azomures*: two weak features of this company are represented by the return on assets and the return on equity. So, the company reported values of these ratios below the value of the inflation rate in 2004 and 2005, in 2006 recorded financial loss, whilst 2007 was the only year when the company presented a level of the return on assets above the level of the inflation rate. Still, the return on equity did not manage to surpass the return on assets, the latter being with 0.32 percentage points lower than the first one. The company capacity to meet the current obligations assumed, reflected by the current liquidity, and had overcome the optimal value in 2004-2007¹⁰. The extent that the company can meet its obligations on long term, represented by the solvability rate, is a good one because this rate has been above the minimum level of 2¹¹ in the period 2005-2007. These two indicators, the current liquidity and the solvability rate, could help to strengthen the investors' confidence in this company. The total debt/total assets ratio reported a negative evolution by recording values higher than the maximum level of 0.5¹² in the reported period. Nevertheless, this situation can be positively interpreted because the borrowed capital had been invested in projects which helped the company to record in 2007 the highest return on assets and return on equity rates reported in the analyzed period. Also, in 2007, the company succeeded to report a profit after a year in which it recorded a financial loss. Besides, the total debt/total assets indicator presents a descending trend, in 2007 it had been very close to the level considered optimal, which strengthens once again the previous statement under which the company doesn't need any more external resources in a high proportion because the necessary investments in order to develop the business have already been processed and they have lead to the increase of the companies' profitability.
- *Biofarm*: in the analyzed period, the current liquidity has been greater than the optimal value, which means a low financial risk of the company. The solvability rate has known a positive trend, which emphasize a decrease of the company debt with almost 25 percent in 2007 related to 2004, while the total assets has increased with 162 percent. The total debt/total assets ratio recorded an ascending evolution and it was below the maximum

value in this period, which means a reduction of the external resources borrowed in order to sustain the investments made by the company.

- *BRD- Groupe Societe Generale*: the current liquidity had values over the optimal level, which means a lower financial risk. The solvability rate had an important fluctuation in 2004-2006, whereas in 2007 became stable at the level of 1.18. This value is sub-optimal, which highlights a growth in firm's leverage. This situation is confirmed by the total debt/total assets rate, which is up to the maximum value in this period of time. This indicator explains the dynamics and the values recorded by the solvability rate, meaning that the company preferred external resources rather than internal ones. From a bank's perspective, this situation is a favourable one because an important percentage of the external resources are represented by the bank deposits, which confirms clients' confidence and its high performances.
- *S.S.I.F. Broker*: the companies' financial risk is very low because, in each year, the current liquidity and the solvability rate were greater than the minimum level. Besides, the total debt/total assets had reported values under the maximum level. Even though these indicators have recorded optimal values, the trend is inadequate: over the years the current liquidity and the solvability rate have decreased and the total debt/total assets have increased.
- *Impact Developer & Contractor*: the current liquidity had a positive trend, with values higher than the minimum level in the period 2004-2007. Instead, in 2006, the solvability rate had a value less than the minimum level, but in 2007 it reached a normal value. A different trend had the ration total debt/total assets, with the highest value in 2006 and the minimum one in 2007.
- *Rompetrol Rafinarie*: the solvability rate presented has grown from 1.08 in 2004, under the minimum value of 2, to levels over the minimum value in 2005 and 2006. In 2007, it reached again a sub-optimal value. Regarding the current liquidity, the company has reported values up to the minimum level in the analyzed period. The leverage exceeded the maximum level in the period 2004-2007, which could signify an unprofitable use of the external resources. This statement is also strengthened by the fact that the company did not succeed to obtain profit in 2006 and 2007.
- *Petrom*: its indicators had acceptable values, although the current liquidity reduced from 4.79 in 2004 to 1.82 in 2007, a situation that could influence negatively the investors' decision. The level of leverage was substandard, while the solvability rate has reported more than satisfactory values.
- *C.N.T.E.E. Transelectrica*: the current liquidity was lower than the minimum value in 2004, but in 2005-2007 it reported a positive change. In the analyzed period, except year 2005, the solvability rate was superior to the minimum level. Although it recorded values greater than the maximum level, the total debt/total assets ratio decreased from 1.65 in year 2005 to 0.68 in 2007. Consequently, we need to focus over these ratios in the following period in order to analyse the companies' ability to pay off the long term debt and its solvability.
- *S.N.T.G.N Transgaz*: the company had some difficulties with its liquidity, with values less than the minimum value for 2005 and 2006. Instead, the company had no problem with its solvability rate because the values were greater than the minimum level for the entire period. The total debt/total assets had been superior to the maximum value, with a decrease in 2007. However, in 2007 this rate is very close to the optimal value. Moreover, in the same

year, the current liquidity had overcome the minimum value, which determined a lower need for the external resources.

- *Transilvania Bank*: regarding the solvability rate in 2004-2006, the company could reach some difficulties in paying off its total debt. Another weak point is leverage, up to the standard (maximum) level in the entire period of time, with an ascending trend. Instead, the company had no difficulties as regards both its liquidity and solvability. Similar to BRD, Transilvania Bank had leverage higher than the optimal value. Also, in 2007, this ratio is greater than for BRD in the same year. If we will study the Balance Sheets for both banks, we will discover that even though Transilvania Bank had a higher leverage, it did not manage to make more bank deposits than BRD. So, in the case of Transilvania Bank, the higher level of this ratio is not an advantage.

5.2. Fundamental analysis through the capital market indicators

In Annex 2 we presented six capital market indicators for each of the listed companies from the BET Index and the values of these indicators for the entire market, for the period 2005-2009.

For the entire period, there is a growing trend for market capitalization in 2005-2007 and a decreasing one in 2008 and 2009, similar with the evolution for the total turnover. The analysis must be made in an unfavourable general, international context, including the case of the Romanian capital market, with "bear" markets since the beginning of 2008. The market capitalization is an important indicator for the investors with risk aversion because it shows the market value of a company¹³. The company with the highest value for the market capitalization in 2009 is Petrom and it is followed, like in the years before, by BRD-Groupe Societe Generale.

PER and P/BV ratio had general descending trends, which can be favourable for the investors because stocks with lower values for both PER and P/BV offer greater dividend yields than the ones with higher values for these indicators. Impact and Petrom are two appropriate examples through PER and dividend yield¹⁴. Thus, even though Impact has recorded in 2007 the highest value for PER, but the dividend yield was zero, while Petrom reported in 2005 an annual value for PER of 106.75 and the dividend yield was only 0.69, and in 2009 PER dropped to 3.53 and dividend yield rose to 9.94.

Generally, the velocity turnover had a descending trend, without significant variations during the period 2004-2007. This relatively stable situation is preferred by the investors with risk aversion because it gives them confidence in those stocks. A stable turnover velocity signifies an important and continuous volume of transactions. This means a minimal fluctuation for the stock prices and, in addition, it allows the investor to make frequent changes in its portfolio. However, in 2008 and 2009, due to the global financial crisis, in Romania the velocity turnover has known an important reduction at the market level as well as for each company from BET index. Certainly, there are a few exceptions like Azomures, BRD, Impact and Transilvania Bank. This means that the investors' interest in these companies' stocks is still high.

In the following paragraphs, we will present the particular situations reported in the capital market indicators, classifying the selected companies on economic specific sectors:

- **The manufacturing industry** (chemical products and substances manufacture): in our study includes Azomures and Biofarm. Because Biofarm became a part of the BET index in March 2009 it is obvious that the company's strength feature is represented by the turnover velocity. While the financial crisis grew deeper, Azomures reported a level of the turnover velocity of 5.56 in 2008 and 4.26 in 2009, while it recorded only a 0.8 value in

2007. Instead, Biofarm presented a continuous decrease of this ratio from 10.42 in 2005 to 2.06 in 2009. In addition, while the companies included in BET index structure had reductions for market capitalization and total turnover ratios in 2008 compared with the previous year, Azomures presented a double increase of the market capitalization and a nearly 14 times increase of the total turnover. Another strong point of this company is represented by PER, which rose from a negative value in 2007 to a positive one in 2008 and 2009, while the PER reported by Biofarm dropped from the maximum of 29.81 reached in 2007 to 5.15 in the current year. None of the two companies had a dividend payout different from zero. Regarding the P/BV ratio, we can see a decline of this ratio for both companies, a trend that is also highlighted by the entire capital market for the entire period.

- **The financial intermediary sector** is represented by BRD, S.S.I.F. Broker and Transilvania Bank. In terms of market capitalization, all three financial institutions had an ascending evolution in 2005-2007. In 2007, they have reported the highest values of market capitalization and total turnover, excepting Transilvania Bank, which presented the highest values of the total turnover in 2005. After 2007, it followed a significant decrease of those ratios, the most important decline being reported by S.S.I.F. Broker with a 91 percent fall of the market capitalization in 2009 compared with 2007 and a 93 percent decrease of the total turnover in the same period.

Regarding the turnover velocity, it had a significant decrease for Transilvania Bank, from 4.56 in 2005 to 1.95 in 2007. After this year, it followed an ascending trend by reaching 4.63 in 2009. Due to face value's consolidation, its stocks were suspended from trading on the capital market between 15th September 2008 and 7th January 2009, which can be considered as a possible explanation for this financial trend¹⁵. So, the company avoided the major reductions that characterized the capital market in October 2008¹⁶ and it tried to slow down the decrease of its own shares price by buying them back in a maximum 5 percent value of its equity¹⁷.

The most significant variation of the turnover velocity had also S.S.I.F. Broker. This company has known a reduction of this ratio from 10.74 in 2005 to 6.87 in 2009. BRD reported insignificant variations of the turnover velocity, although it dropped from 0.94 in 2005 to 0.49 in 2006, but in the following years it presented increasing values. So, in 2009 it succeeded to slightly overcome the value reached in 2005 by reporting a value of 0.97. Because of the low variation of this ratio, BRD has an advantage compared with the two other companies in this field from the point of view of an investor with risk aversion. Besides, it reported the highest increase of dividend yield, from 2.12 in 2005 to 11.4 in 2009. This evolution can be explained by the companies' desire to maintain the investors' interest in its stocks, especially during the financial crisis, when the dividend yield increased from 1.41 in 2007 to 3.58 in 2008 and 11.4 in 2009. Another explanation could also be the fact that the dividend yield growth is based on the decline of the stocks price, which is a representative situation in the case of a worldwide financial crisis.

The other two companies had not developed a dividend policy. The dividend payout presents a major concern for the investors with risk aversion, who take decisions on long term and they give an important attention to the income based on dividends distribution.

Similar to the trend reported by the capital market regarding PERs' evolution, the financial institutions that we analyzed presented significant reductions of this ratio. The most important variation had in the case of Transilvania Bank, from a maximum level of 37.49

reached in 2006 to 1.31 in 2009, while BRD has known the lowest decrease (from the maximum 25.20 in 2007 to 2.95 in this year).

- **The energetic field** is represented by companies like Rompetrol, Petrom, Transelectrica and Transgaz. This sector of activity draws investors' attention through the significant variation for capital market indicators. So, it recorded notable decrease for PER, whom reported a fall from a positive value in 2005, 65.87, to a negative one in 2008 and 2009, respectively (-26.27) and (-1.19), in the case of Rompetrol. For Petrom, this ratio decreased from 106.75 in 2005 to 3.53 in the current year. Reported to the other two companies, Transelectrica presented the lowest decrease of this ratio, from the maximum value of 19.32 in 2008 to 4.64 in the current year. Moreover, Rompetrol presents major fluctuations for the turnover velocity and the P/BV ratio and, in addition, it had distributed no dividend. By the contrary, Petrom reported an increase of the dividend yield from 0.69 in 2005 to 5.45 in 2008 and 9.94 in 2009, with a decrease in turnover velocity from 2.29 in 2006 (after it has grown with 1.82 in the previous year) to 0.25 in 2008. Nevertheless, we can see a slight increase of the velocity turnover in the case of both companies in 2009, with 0.05 percentage points in the case of Petrom and with 0.42 percentage points in Rompetrol case. Instead, Transelectrica presented a descending trend in the same period.

In the case of Transgaz, it is too soon to draw some conclusions mostly because this company has been listed on the Romanian Stock Exchange in January 2008. However, our attention is drawn by the fact that the P/BV ratio recorded an upper level in 2008, compared with other three analyzed companies in the same year. Also, the values of PER and the dividend policy represent an advantage against Transelectrica and Rompetrol. Thus, if in the case of Transelectrica, PER presented a 14.68 decrease in 2008-2009, in Transgaz case, it has reduced with 4.47 in the same period, while Rompetrol presented a negative value, as we mentioned earlier. A solid point is represented by the dividend yield, which has increased from 4.36 in 2008 to 6.24 in 2009, while Transelectrica has known an insignificant rise of this indicator with only 0.17 percentage points in the same period and Rompetrol did not develop a dividend policy in the entire period.

According to the turnover velocity, Transgaz presented in 2009 the lowest value and it is followed by Petrom. However, these two companies remain an attractive investment for investors with risk aversion due to the higher levels of the dividend yield recorded.

- In the **construction sector** the only company included in the BET Index is Impact and this is the reason why the specific values for the capital market indicators will be compared with the ones recorded by the capital market, as a whole. This company had significant variations for PER, which rose from 17.46 in 2005 to 77.16 in 2007 and after that it followed a decrease to 33.46 in 2008 and continued up to 4.67 in 2009. The high values of this ratio can be explained through its overvaluation, which became undervalued in 2009. Although, the market value of PER had a similar trend, with low fluctuations, from 24.43 in 2005 to 3.26 in 2009. The turnover velocity had an ascending evolution from 1.9 in 2006 to 5 in the current year, while the same indicator reported by the market has known a decline from 14.03 to 9.56, in the same period. Although, the P/BV ratio emphasizes a face value greater than the market value, up to 0.13 in 2009, which emphasizes an unfavourable feature for the company image. Moreover, the company had no dividend payout. It is necessary to mention the fact that the changes occurred in the case of the turnover velocity and the P/BV ratio are due to the consolidation procedure developed from September 2008 to February 2009¹⁸ and to the process of buying back 4

percent of its own stocks¹⁹. These measures determined the stocks price's stabilization, which were affected by the real estate' crisis which followed the financial crisis. The real estate' crisis developed especially in the fourth trimester of 2008.

6. Score method for choosing a stock portfolio for a private administrated pension fund

The private administrated pension funds are those entities that will provide, in Romania, in the upcoming years, the private pension system alongside the public pension. The level of pension will depend on the net asset value created and managed by these funds. In this context, the assets will depend significantly on the efficiency of investments in various financial instruments and the final decision, in this context, belong to the fund's administrator. In this article, for the selection of listed stocks to be included in a portfolio, we propose **a score function**. Using this method we determine which of the companies analyzed previously had the best performance in terms of an investor with risk aversion, and the final goal will be identify the best three companies included in BET index in terms of return and risk.

The proposed "score function" will have the following form:

$$Z = \sum_{i=1}^n p_i \cdot R_i$$

where:

p_i – the relative weight associated to the indicator (rapport) R_i ;

R_i – the rapport taken into consideration;

n – number of the significant rapports for the model.

Depending on the outcome raised from the function score, it will be identified the most attractive stocks from a minority shareholder's point of view, with risk aversion, investing capital market in Romania.

The significant indicators (rapports) to be included in the score function derived on the above indicators analyzed in this article, taking into account the findings of a survey included in the study of Dragota and Serbanescu (2009) about **the behaviour of investors on the capital market in Romania**²⁰.

Through this study, were obtained information on the investment options of the participants on the stock exchange, their attitude towards risk, how to protect against this risk, their information sources and the frequency at which investors appeal to those data, the foresight about their portfolio's yields, and a series of demographic information on employment status and position occupied, the level and nature studies, their age and "maturity" on the Romanian capital market. Respondents received either directly or by e-mail a questionnaire including 17 questions. The questionnaire was sent to the Romanian individual investors²¹.

The study of Dragota and Serbanescu (2009) describe the **selection criteria** taking into consideration by the investors for their investment decision as regards the stocks included in the portfolio. The investors participating in the survey considered that the main criterion to be considered is profitability, 64.86% of them giving it maximum points. The second most important criterion, according to respondents, is liquidity, 51% of respondents giving to liquidity a score of 4/5. On the third place was considered the risk. The previous performances and industry are considered less important by the respondents. In this ranking, it must be identified the switch between liquidity (most important) and risk, which may be a clue to the speculative' behaviour of

the investors on the Romanian capital market. It must be point out that the study of Vasilescu and Vatui (2004 a, b, c) identified as benchmark for the investment decision the liquidity of securities.

Regarding **the perception of risk**, approximately 80% of respondents define risk as the variability of the market price. Only 19% believe that the risk means the bankruptcy risk, while 17% associated the risk with the possibility of no cash dividend. From the total number of respondents, 7.43% define risk as the possibility that the securities to be suspended from the transaction on the stock market. Only one respondent consider nationalization as a possible risk, which can be viewed as a confidence of investors in terms of economic development of Romania as a market economy.

The share π_i associated with each score R_i recommended to be used to develop stocks portfolio will be based on these findings, the method used for processing the data being the **Likert's scale**. **The Likert Scale** is a rating device frequently used in marketing research questionnaires in which respondents indicated their level of agreement with a statement by choosing the appropriate response from the scale, e.g. strongly disagree, disagree, undecided, agree, strongly agree.

In the analysis conducted up to this moment for the listed stocks included in BET index, we select those indicators (R_i) that correspond to the preferences of the Romanian investors who have been surveys, namely:

1. the dividend yield (capital market indicator) has a special importance in our score function because the minority shareholders are concerned, firstly, about the dividends and then about the capital gains. In addition, "dividends are seen as an object of consumption, while the shares as an investment, hence the reluctant to sell shares and encouraging earnings from dividends (Dragota V. **Dividend policy. An approach in the context of economic environment from Romania**, 2003);
2. the total turnover (capital market indicator) because the investor is interested to conduct frequent trading in its portfolio to avoid losses from changing in market prices. In addition, a high rate of liquidity guarantees a certain stability in the stock price;
3. the current liquidity ratio (financial indicator) because, as we noted above, minority shareholders are investors with risk aversion;
4. total debt/total assets (financial indicator) have less interest for this class of investors. However, its importance arises from the fact that its grater value shows the volatility of the net income and the increase of the insolvency risk. Since the current liquidity ratio is considered an indicator to quantify risk, we use both indicators in the score function, using their arithmetic average;
5. PER reflects the price that investors are willing to pay for a monetary unit of net income per share or the number of years needed to recover the investment, when the full distribution of net profit as dividends. In general, it is estimated that stocks with a PER value above the average value for the industry are overvalued, while those with a PER value under the average value for the industry are undervalued. However, the trend of PER must be compared with the evolution of the financial indicators, taking into consideration that this coefficient is influenced by the accounting policy used by the company. It is therefore considered that a normal level of PER is within the range 15-25²².

In the score function, based on the findings from the cited survey, we will consider a rate between each company's PER and the average value of PER for the economic sector in which it belongs.

We calculate the weights related to each indicator considered in the analysis, using Likert scale and the results of the survey carried out among investors on the Romanian capital market, as follows:

$$a) \quad p_{profitability} = \frac{1x0 + 2x6 + 3x12 + 4x34 + 5x96}{148} = 4.48 ;$$

$$b) \quad p_{liquidity} = \frac{1x0 + 2x11 + 3x29 + 4x75 + 5x33}{148} = 3.88 ;$$

$$c) \quad p_{risk} = \frac{1x6 + 2x19 + 3x11 + 4x6 + 5x6}{148} = 2.91 ;$$

$$d) \quad p_{previous_performances} = \frac{1x25 + 2x104 + 3x13 + 4x6 + 5x0}{148} = 2.00 ;$$

$$e) \quad p_{sector} = \frac{1x128 + 2x8 + 3x7 + 4x5 + 5x0}{148} = 1.25 .$$

The score function will be the following:

$$Z = 4.48 X_1 + 3.88 X_2 + 2.91 X_3 + 2 X_4 + 1.25 X_5$$

where:

X_1 - the current dividend yield;

X_2 = the total turnover;

$$X_3 = \frac{\text{current_liquidity_ratio} + \text{total_debt}/\text{total_assets}}{2} ;$$

X_4 = the dividend yield from the previous year;

$$X_5 = \frac{PER_{company}}{PER_{sector}} .$$

Table 2. The application of the score method for the stocks included in BET index

Company	X_1	X_2^*	X_3^{**}	X_4	X_5^*	Z
Azomures S.A.	0	2,52	1,06	0	4,37	18,31
Biofarm S.A.	0	4,98	1,82	0	-2,42	22,15
BRD - Groupe Societe Generale S.A.	11,4	0,7	5,92	3,58	1,06	79,48
S.S.I.F. Broker S.A.	0	8,06	3,16	0	0,76	41,42
Impact Developer & Contractor S.A.	0	3,19	2,63	0,16	3,32	24,51
Rompetrol Rafinare S.A.	0	2,3	1,2	0	1,6	14,12
Petrom S.A.	9,94	0,79	1,65	5,45	1,91	65,70
C.N.T.E.E. Transelectrica	3,78	1,03	1,03	3,61	2,69	34,51
S.N.T.G.N. Transgaz S.A.	6,24	0,43	0,89	4,36	2,33	43,85
Banca Transilvania S.A.	0	3,36	7,04	0	2,35	36,45

* were taking into consideration the average values for the period January 2005 – March 2009;

** were taking into consideration the average values for the period 2004-2007.

According to the results obtained, companies have the best performance are BRD - Groupe Societe Generale, Petrom and SNTGN Transgaz. Except BRD, which gave shareholders dividends in all years examined, Petrom has distributed dividends only in the years 2006-2008, while Transgaz distributed dividends every year but we must take into account the fact that it was listed on BSE in early 2008. It also shows that the three companies have shown no considerable variation in the total turnover, but experienced significant fluctuations in market capitalization,

for the coefficient PER, similar to changes from the capital market, as a whole, in the reported period. Instead, BRD offset the significant variability for the PER coefficient offset by obtaining the highest dividend yield for the three selected companies, for the entire period.

7. Conclusions

The study was designed from the perspective of an investor with risk aversion as should be considered the private pension funds. Accordingly, we considered as representative for the fundamental analysis the most ten liquid shares listed on the Bucharest Stock Exchange, included in the BET index. The main indicators used to measure the companies' performances were the dividend yield, total turnover, PER, current liquidity ratio and the leverage. The three stocks considered representative for an investor with risk aversion were selected based on a score function, the weights of each indicator being the result of the use of Likert scale. Since no model was built for the economic environment in Romania, we considered a model that had the coefficients for the score function both financial and capital market indicators. Thus, the best companies in terms of an investor with risk aversion have proved to be BRD - Groupe Societe Generale, Petrom and SNTGN Transgaz.

At this moment, the transactions from Bucharest Stock Exchange are purely speculative. This is due to the international financial crisis generated by sub prime mortgage crisis with upper risk from the United States of America and by the turbulence in international financial markets, with significant effects on the Bucharest Stock Exchange, too. Therefore, it is recommended that investment decisions on the capital market in Romania should be taken with caution.

References

1. Alexander, D. and Britton, A. **International financial reporting and analysis**, Thomson Printing House, Padstow, 2003
2. Anghelache, G. **Piata de capital din Romania. Caracteristici, evolutii, tranzactii**, Economica Printing House, Bucharest, 2004
3. Dragota, M. (coord.), Dragota, V., Handoreanu, C., Stoian A., Serbanescu, I.C. and Obreja Brasoveanu, L. **Piete financiare: structura, institutii, instrumente, reglementari**, ASE Printing House, Bucharest, 2009
4. Dragota, V. and Serbanescu, V. **Cateva indicii privind comportamentul investitorilor de pe piata de capital din Romania. Rezultatele unei anchete**, Theoretical and Applied Economics Journal, Bucharest, 2009
5. Dragota, V. **Politica de dividend. O abordare in contextul mediului economic din Romania**, ALL BECK Printing House, Bucharest, 2003
6. Dragota, V., Ciobanu A., Obreja L. and Dragota M. **Management financiar. Vol 1: Analiza financiara si gestiune financiara operationala**, Economica Printing House, Bucharest, 2003
7. Dragota, V., Dragota, M., Damian, O. and Mitrica, E. **Gestiunea portofoliului de valori mobiliare**, Economica Printing House, Bucharest, 2009
8. Halpern, P., Weston J. F. and Brigham, E. F. **Finante manageriale**, Economica Printing House, Bucharest, 1998
9. Roman, M. **Statistica financiar-bancara si bursiera**, ASE Printing House, Bucharest, 2003
10. Sharpe, W. F. and Alexander, G. J. **Investments**, Prentice Hall Printing House, Upper Saddle River, 1999
11. Stancu, I. **Finante**, Economica Printing House, Bucharest, 4th edition, 2007.
12. Serbanescu, C. **Asigurari si protectie sociala. Tendinte europene**, Universitara Printing House, Bucharest, 2008
13. Valceanu, Gh., Robu, V. and Georgescu, N. **Analiza economico-financiara**, Economica Printing House, Bucharest, 2005

14. Vasilescu, C. and Vatui, M. **Informatiile financiar-contabile si bursiere pe piata de capital romaneasca**, Finante, Banci, Asigurari, No. 7 (79), year VII, July 2004, pp. 28-34, [2004 c]
15. Vasilescu, C. and Vatui, M. **Investitorii persoane fizice si comportamentul lor investitional pe piata de capital romaneasca**, Finante, Banci, Asigurari, No. 3 (75), year VII, March 2004, pp. 31-33, [2004 a]
16. Vasilescu, C. and Vatui, M. **Puterea financiara si preferintele investitorilor persoane fizice pe piata de capital romaneasca**, Finante, Banci, Asigurari, No. 4 (76), year VII, April 2004, pp. 17-22, [2004 b]
17. * * * Law 297/2004 regarding the capital market
18. * * * Law 411/2004 regarding privately administrated pension funds
19. www.bvb.ro
20. www.dailybusiness.ro
21. www.csspp.ro
22. www.cnb.cz
23. www.curierulnational.ro
24. www.zf.ro
25. www.ktd.ro
26. www.ssifbroker.ro

Annex 1.

Analysis of economic and financial situation for the companies included in the BET Index²³

Company	Indicator	2004	2005	2006	2007
The inflation rate		11,9%	9%	4,87%	6,7%
Azomures	Return on assets	10,63%	6,46%	0	16,65%
	Return on equity	11,36%	6,46%	0	16,33%
	Current liquidity	1,17	1,22	1,2	1,64
	Solvability rate	1,85	2,16	2,23	3,02
	Total debt/Total assets	1,11	0,82	0,77	0,51
Biofarm	Return on assets	16,75%	19,32%	17,1%	8,75%
	Return on equity	19,03%	20,45%	18,03%	8,95%
	Current liquidity	2,29	2,57	3,88	4,8
	Solvability rate	3,59	4,19	6,30	12,63
	Total debt/Total assets	0,39	0,31	0,19	0,09
BRD - Groupe Societe Generale	Return on assets	2,47%	3,92%	3,32%	3,77%
	Return on equity	24,31%	22,71%	21,95%	26,53%
	Current liquidity	3,70	3,57	3,27	2,55
	Solvability rate	1,24	4,19	1,18	1,18
	Total debt/Total assets	6,62	7,22	10,7	9,66
S.S.I.F Broker	Return on assets	18,18%	28,18%	22,56%	24,28%
	Return on equity	18,18%	28,19%	22,63%	24,33%
	Current liquidity	11	5,69	4,68	3,25
	Solvability rate	14,69	7,35	4,88	7,44
	Total debt/Total assets	0,07	0,16	0,26	0,16
Impact Developer & Contractor	Return on assets	13,64%	4,51%	5,52%	1,66%
	Return on equity	19,07%	6,98%	10,06%	2,27%
	Current liquidity	2,54	4,25	6,04	4,45
	Solvability rate	2,35	2,11	1,84	2,49
	Total debt/Total assets	0,82	0,98	1,27	0,69
Rompetro Rafinare	Return on assets	1,23%	7,1%	0%	0%
	Return on equity	10,93%	18,03%	0%	0%
	Current liquidity	1,66	1,53	1,33	1,06
	Solvability rate	1,08	2,56	2,46	1,93
	Total debt/Total assets	1,18	0,94	0,77	1,07
Petrom	Return on assets	0%	9,13%	13,36%	9,62%
	Return on equity	0%	13,2%	12,66%	13,49%
	Current liquidity	4,79	3,2	2,78	1,82
	Solvability rate	6,53	7,83	8,50	8,11

Company	Indicator	2004	2005	2006	2007
C.N.T.E.E. Transelectrica	Total debt/Total assets	0,13	0,15	0,13	0,15
	Return on assets	2,34%	1,92%	8,57%	1,47%
	Return on equity	2,51%	4,37%	12,61%	2,16%
	Current liquidity	0,24	1,11	1,45	1,44
	Solvability rate	2,11	1,67	2,52	2,58
S.N.T.G.N. Transgaz	Total debt/Total assets	0,97	1,65	0,7	0,68
	Return on assets	-	7,87%	18,79%	9,75%
	Return on equity	-	13,4%	19,85%	14,30%
	Current liquidity	-	0,89	0,3	1,84
	Solvability rate	-	2,11	2,38	2,84
Transilvania Bank	Total debt/Total assets	-	0,94	0,77	0,58
	Return on assets	2,97%	2,73%	1,94%	2,47%
	Return on equity	20,16%	21,52%	16,84%	42,93%
	Current liquidity	4,50	3,73	4,16	4,41
	Solvability rate	1,14	1,14	1,15	1,15
	Total debt/Total assets	7,54	9,50	10,16	12,26

Annex 2
Capital market indicators for companies included in the BET Index for
the period 2005-2009²⁴

Company	Year	Market capitalisation (RON)	Total turnover (RON)	PER	DIVY	Turnover velocity	P/BV
Capital market indicators	2005	56.917.130.000	6.871.900.000	24,43	0,92	19,6	3,33
	2006	73.341.790.000	9.725.890.000	18,03	1,72	14,03	2,72
	2007	85.962.390.000	13.512.880.000	19,37	2,16	16,05	3,05
	2008	45.701.490.000	6.831.840.000	4,11	8,57	4,75	0,76
	2009	38.453.220.000	679.620.000	3,26	10,61	9,56	0,61
Azomures	2005	150.708.349	2.152.661,27	4,83	3,5	1,35	2,87
	2006	102.269.511	668.023,16	9,77	0	0,63	0,39
	2007	100.384.560	847.805,63	-33,02	0	0,8	0,38
	2008	202.654.071	12.453.937,56	2,88	0	5,56	0,62
	2009	100.472.233	4.283.823,67	0,39	0	4,26	0,29
Biofarm	2005	194.007.361	19.358.392	22,34	0	10,42	-
	2006	249.151.902	14.039.961	23,54	0	5,86	4,72
	2007	390.847,733	14.198.089,33	29,81	0	3,7	5,81
	2008	257.403.241	31.109.481,1	18,48	23,3	2,9	2,48
	2009	73.209.739	1.577.729	5,15	0	2,06	0,48
BRD - Groupe Societe Generale	2005	7.572.995.536	44.660.266,85	23,61	2,12	0,94	-
	2006	12.340.964.381	61.017.780,42	24,98	1,67	0,49	6,54
	2007	17.486.420.589	89.869.123,92	25,20	1,41	0,51	7,89
	2008	11.240.440.734	59.366.521,03	12	3,58	0,58	4,69
	2009	3.751.653.171	36.626.310	2,95	11,4	0,97	1,09
S.S.I.F. Broker PLC	2005	80.331.251	9.308.776,4	28,89	0,16	10,74	-
	2006	119.204.537	8.360.801,08	12,88	0,92	6,56	2,94
	2007	302.775.530	27.399.009,42	13,71	0	8,45	4,5
	2008	173.446.942	10.920.444,58	5,33	0	7,68	1,63
	2009	24.836.670	1.861.463	2,77	0	6,87	0,16
Impact Developer & Contractor	2005	306.459.433	9.836.652,58	17,46	2,22	2,89	-
	2006	485.166.753	9.204.108,83	51,77	0,35	1,9	2,82
	2007	924.861.932	19.722.874,92	77,16	0	2,09	5,82
	2008	348.533.333	9.106.916,02	31,36	0,16	4,07	1,5
	2009	43.600.000	1.560.809,33	4,67	0	5,00	0,13
Rompertol Rafinarie	2005	1.872.456.621	84.703.526,42	65,87	0	4,04	-
	2006	1.892.429.230	45.553.791,75	55,86	0	2,45	2,38
	2007	2.040.835.742	35.724.815,08	-18,04	0	1,69	1,08

Company	Year	Market capitalisation (RON)	Total turnover (RON)	PER	DIVY	Turnover velocity	P/BV
	2008	1.001.566.680	12.726.488,37	-26,27	0	1,26	0,54
	2009	548.581.175	9.237.387	-1,19	0	1,68	0,27
Petrom	2005	19.197.918.837	48.893.270,33	106,75	0,69	0,47	-
	2006	8.477.639.783	27.178.442,92	25,35	2,22	2,29	1,76
	2007	28.175.297.377	67.301.371,75	17,85	2,98	0,67	2,8
	2008	20.840.311.525	4.396.6574,7	10,62	5,45	0,25	1,61
	2009	8.893.125.008	26.754.542,95	3,53	9,94	0,30	0,68
C.N.T.E.E. Transelectrica	2006	2.222.391.297	3.090.5639,2	18,46	0,87	1,44	2,46
	2007	2.885.767.193	25.430.911,08	16,42	2,87	0,93	1,57
	2008	1.578.460.991	15.154.113,98	19,32	3,61	0,94	0,69
	2009	698.823.286	5.651.926,33	4,64	3,78	0,8	0,3
S.N.T.G.N. Transgaz	2008	2.277.717.134	13.907.760,9	9,91	4,36	0,57	2,52
	2009	1.259.408.846	3.680.913,33	5,44	6,24	0,29	0,8
Transilvania Bank	2005	1.902.935.155	86.264.698,17	33,51	0	4,56	-
	2006	3.334.077.644	77.001.484,75	37,49	0	2,58	6,84
	2007	4.132.479.049	77.916.597,33	30,94	0,29	1,95	6,41
	2008	3.243.420.181	61.791.751,81	13,8	0	3,09	4,81
	2009	787.707.495	37.359.092,67	1,31	0	4,63	0,65

¹ Acknowledgements

This research was supported by the Romanian Ministry of Education and Research – the National Authority for Scientific Research (NASR) through National Programme Ideas (PN II), Grant No. 1831/2008.

² www.csspp.ro

³ Serbanescu, C. **Asigurari si protectie sociala. Tendinte europene**, Universitara Printing House, Bucharest, 2008

⁴ www.cnb.cz

⁵ Law no. 411/2004 regarding privately administrated pension funds.

⁶ www.csspp.ro

⁷ Dragota, M. (coordinator), Dragota, V.; Handoreanu, C.; Stoian A.; Serbanescu I. C. and Obreja Brasoveanu, L. **Piete financiare: structura, institutii, instrumente, reglementari**, ASE Printing House, Bucharest, 2009

⁸ Anghelache, G. **Piata de capital din Romania. Caracteristici, evolutii, tranzactii**, Economica Printing House, Bucharest, 2004

⁹ Stancu, I. **Finante**, Economica Publishing House, 4th edition, Bucharest, 2007

¹⁰ Valceanu, Gh., Robu, V. and Georgescu, N. **Analiza economico-financiara**, Economica Publishing House, Bucharest, 2005

¹¹ Stancu, I. **op.cit**, 2007

¹² Dragota, V., Ciobanu, A., Obreja, L. and Dragota, M., **Management financiar. Vol 1: Analiza financiara si gestiune financiara operationala**, Economica Publishing House, Bucharest, 2003

¹³ Stancu, I., **op.cit**, 2007

¹⁴ Stancu, I. **op.cit**, 2007

¹⁵ The official Statements from Transilvania Bank, www.bvb.ro

¹⁶ Daily Business, **Cat de ferite de criza sunt actiunile Bancii Transilvania?**, 29.10.2008, www.dailybusiness.ro

¹⁷ Chirileasa, A. **Banca Transilvania si Impact rascumpara actiuni proprii ca sa le opreasca din scadere**, 28.07.2008, www.zf.ro; Curierul Național Editorial Staff, **Banca Transilvania a rascumparat si vineri actiuni proprii, in valoare de 151.000 lei**, 13.01.2009, www.curierulnational.ro

¹⁸ The official Statements from Impact Developer & Contractor S.A., www.bvb.ro

¹⁹ Ziarul Financiar Editorial Staff, **Impact rascumpara inca un pachet de 550 mii de actiuni**, 12.08.2008, www.zf.ro

²⁰ Dragota, V. and Serbanescu, V., **Cateva indicii privind comportamentul investitorilor de pe piata de capital din Romania. Rezultatele unei anchete**, Theoretical and Applied Economics Journal, Bucharest, 2009

²¹ Most respondents (99.32%) invest in stocks, the most common financial instrument on Bucharest Stock Exchange (in fact, only one respondent had no shares in his portfolio). It can be noted, however, the interest for portfolio diversification, being considered the investments in mutual funds (78.38% of respondents), bank deposits (70.27%), derivatives (63.51%) and bonds (25.68 %) (Dragota and Serbanescu, 2009).

²² Valceanu, Gh., Robu, V. and Georgescu, N. **Analiza economico-financiara**, Economica Printing House, Bucharest, 2005.

²³ The financial indicators were determined based on the Balance Sheets for the listed companies, from www.bvb.ro.

²⁴ The capital market indicators were determined based on the Monthly Reports published on the BSE website, www.bvb.ro.

SPECTRALYZER: A COMPREHENSIVE PROGRAM TO CLASSIFY FTIR MICROSCOPIC DATA APPLIED FOR EARLY DETECTION OF CRITICAL AILMENTS

Jeremy R. SCHWARTZ¹

Research Assistant, Negev Monte Carlo Research Center (NMCRC) and
Department of Software Engineering, Shamoon College of Engineering, Beer Sheva, Israel

E-mail: jrschwar@gmail.com

Shlomo MARK^{2,3}

PhD, Senior Lecturer, Negev Monte Carlo Research Center (NMCRC) and
Department of Software Engineering, Shamoon College of Engineering, Beer Sheva, Israel

E-mail: MarkS@sce.ac.il

I. YAAR

Nuclear Research Center Negev, Beer-Sheva, Israel

E-mail:

S. MORDECHAI

Department of Physics and the Cancer Research Center,
Ben-Gurion University, Beer-Sheva, Israel

E-mail:

Abstract: *Micro spectroscopy can be used for the early diagnosis of critical ailments. Early diagnosis is crucial for identifying the presence of a disease before it progresses beyond a critical stage. It has been shown that micro spectroscopy can be used as a non-invasive or limited-invasive approach for detecting different stages of cervical intraepithelial neoplasia.. The unique spectral "fingerprint" that characterizes premalignant cells can be used to differentiate each stage from normal (healthy) cells. Techniques lacking automatic, objective, sensitive and rapid diagnostic tools are not sufficient. We have developed SPECTRALYZER, a novel micro spectroscopical computational tool for automated complementary diagnostic analysis.*

Key words: *Micro spectroscopy; computational tool; SPECTRALYZER*

1. Background

Cervical cancer especially invasive squamous cell carcinoma is the second most prevalent cancer among women [1-4]⁴ and squamous cell carcinoma constitutes 80–90% of cervical cancers. Screening programs, especially the identification of precancerous lesions at an early stage, can produce good prognoses and prompt early treatment to prevent advanced-stage cancer and death [5]

Current tests, such as PAP smears and a polymerase chain reaction [6], though they can detect cervical cancer, do not give the total diagnosis unless they are supported by histology or results of biopsies.

During the onset of cervical cancer and neoplasia, the biochemical composition of tissues becomes altered [7]. Several approaches to quantitatively link these changes with spectral characteristics [8, 9] involve sophisticated cell mechanisms and techniques to interpret them. Fourier-transform infrared micro spectroscopy (FTIR-MSP) has been shown to be a promising diagnostic tool in monitoring biochemical changes in cells during experiments with exfoliated cervical cells and biopsies [10]. Materials with differing chemical compositions or structures will exhibit different absorption spectra, which can be used as "fingerprints" for the characterization of these materials [11], and thus the biochemical changes between neoplastic stages can be used to identify them.

2. Program

THE SPECTRALYZER was developed in order to automate and facilitate the Fourier-Transform Infrared (FTIR) analysis on biological samples. Many spectroscopic tools can generate a text-based list of IR spectral data, i.e. coordinate points representing absorbance as a function of wavenumber, that can be saved as simple text files Investigation Spectroscopy in the mid-IR spectral range includes analysis and comparison between many samples and between the sample and known markers to identify the unique characteristics of the sample which might reveal the unique substance found in one sample but not the other. The idea beyond the developing of the SPECTRALYZER is to build an automatic, objective, sensitive and rapid analyze tool that would designated for early detection of biological spectroscopy. The SPECTRALYZER can easily read multiple spectra files for fast and accurate scrutiny and comparison. Spectra may be manipulated and analyzed by this program in many different ways: subtracted, shifted, scaled, and "smoothed", as well as matching peaks, principal component, and linear discriminant analysis.

A relational database development was required to process this vast information. Such a database allows further analyses to discover new substances and to collect biological knowledge of the studied sample(s). This task required the use of several bioinformatics software tools and database searches, executed in a batch mode and often required additional programming.

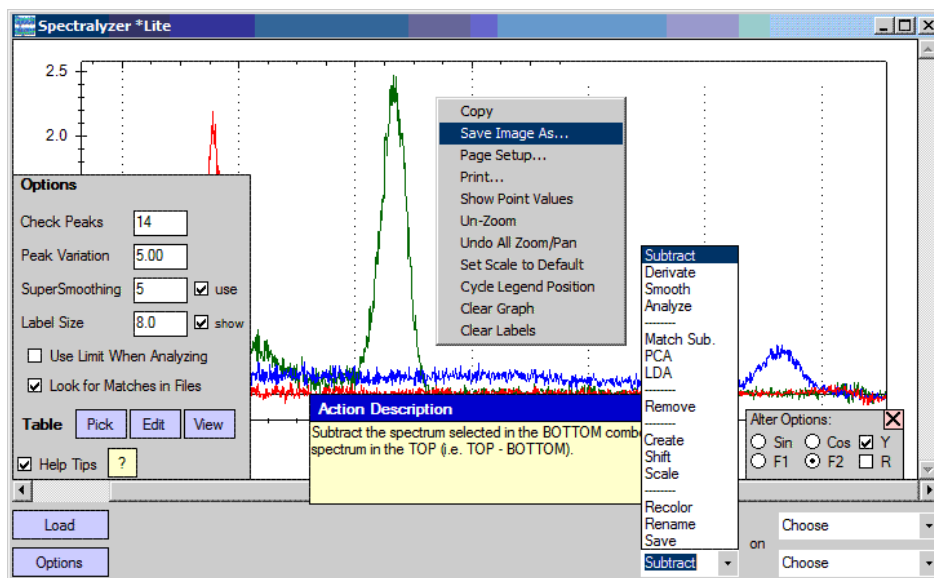


Figure 1. SPECTRALYZER main window with sample spectra and Options, Alter, and Right-Click menus shown

The program window is divided into two main sections -- the graphical pane and the toolbar (Figure 1.). The toolbar, which contains action buttons and selection boxes, is located along the bottom of the window to allocate the majority of the window area for the graphical pane located above. There are two selection boxes (combo boxes) for choosing the spectrum and the action to be performed. Buttons will load spectra, show the options menu, or perform the selected action.

Two additional information panes, normally hidden, will let users view or change some of the program settings or alter the selected spectrum according to the current action, respectively. There are also popup dialogs reporting useful information about each button or selected action.

3. Technical Background and Software Development

The SPECTRALYZER was then constructed in C#, using Microsoft Visual Studio .NET 2003 [12]. The SPECTRALYZER program was developed under the requirements and conditions of Object Oriented Analysis Design and Programming [13- 16] with particular paid attention to form techniques such as UML (Unified Modeling Language) diagramming [17-19] and design patterns [20, 21]. The use of iterative design, parallel development, and high modularity all enhance maintainability, ease of verification, and code reusability.

The SPECTRALYZER program has been developed as a two iteration process of three steps each: alpha, beta, and release. The first iteration was a module serving as an independent confirmation tool. After a Software Test Plan, Description, and Report according to MIL-STD-498 [22], the program was moved to C# under Visual Studio .NET 2005 in the next iteration in order to more easily build a user interface. At the second iteration the SPECTRALYZER passed the alpha stage; having undergone the major testing phases (requirements, design, program, and installation) [23, 24], it was reviewed under rigorous real-life conditions by typical users in the beta step. User feedback, suggestions, and accuracy assessments were incorporated into the program, which was subjected to

regression testing to make sure none of the improvements cause earlier passed tests to fail. After success it was published as the release version.

4. Advantages of SPECTRALYZER

The SPECTRALYZER program contains many of the same abilities as commercial products like Bruker Optic's OPUS, such as easy spectrum manipulation, saving spectral graphs for printing, and spectra library management, yet it comes in a package hundreds of times smaller. SPECTRALYZER was created specifically for analyzing photonic spectra, especially those from the FTIR spectroscopy of biological samples. It is not enough to use one method to find the markers for various ailments. This program has the ability to readily use multiple methods for analyzing different spectra, improving the chances of accurately detecting the important markers.

These methods include spectrum comparison, principal-component analysis (PCA) [25], and linear-discriminant analysis (LDA) [26, 27]. For discovering component elements (represented by specific peaks in a spectrum), the user can build, arrange, and update their own database which is used for peak matching. Scaling a spectrum by a ratio, such as DNA/RNA ratio ($1121/1020\text{ cm}^{-1}$) [28], is a simple action. All of the functions of the program are tightly integrated into a straightforward and consistent interface. Many of the default or underlying system options are easily changed in the settings file, and the entire program is ready for localization to most languages just by changing a single text file -- both of which are fully explained within their respective files. The entire program requires only a few files to begin, and is ready to run on a Windows with .NET computer without any installation -- thus the program could be loaded on a disk and transferred between workstations anywhere. The SPECTRALYZER, implement a "protocol macro" function, whereby automated multiple actions may be combined and saved for rapid and easier repetition.

6. Sample preparation

The method of Argov et al. [29] was followed for sample preparation. Biopsy samples were taken from several stages of neoplastic cells -- "normal" (healthy) cells, "CIN1" stage is mildly dysplastic, moderately dysplastic stage "CIN2", a severe dysplasia is "CIN3" cases (which is considered as cancer *in situ*), and an "invasive" stage that has progressed beyond CIN3.

Two adjacent paraffin sections were cut from each biopsy; one was placed on a zinc-selenium slide and the other on glass slide. Care was taken to ensure that tissue sections were practically identical with a thickness of $10\text{ }\mu\text{m}$. The first slide was deparaffinized using xylol and alcohol and was used for FTIR measurements; the second slide was stained with hematoxylin and eosin for parallel histology review. In cases where the biopsy showed histological similarity to more than one neoplastic stage, the different regions were identified and measured separately with the help of an expert pathologist who distinguished the stages under microscope.

The microscopic sites of measurements were taken keeping in view the occurrence and diagnostic features of the neoplasia in relation to its staging as described [30,31] To achieve high signal to noise ratio (SNR) 256 co-added scans were collected in each

measurement in the wavenumber region 600 to 4000 cm^{-1} . The spectra were baseline corrected automatically by the sensing equipment using OPUS software. The spectra were normalized to the amide I (1652 cm^{-1}) absorbance peak for calculations and subsequent analyses.

7. Analysis demonstration

An FTIR spectrum was selected from each cervical cancer stage to demonstrate how the SPECTRALYZER program may be used. Opening the program immediately requests a spectrum file to load, and selecting the cervical spectrum file loads all five distinctive spectra as seen in Figure 2.

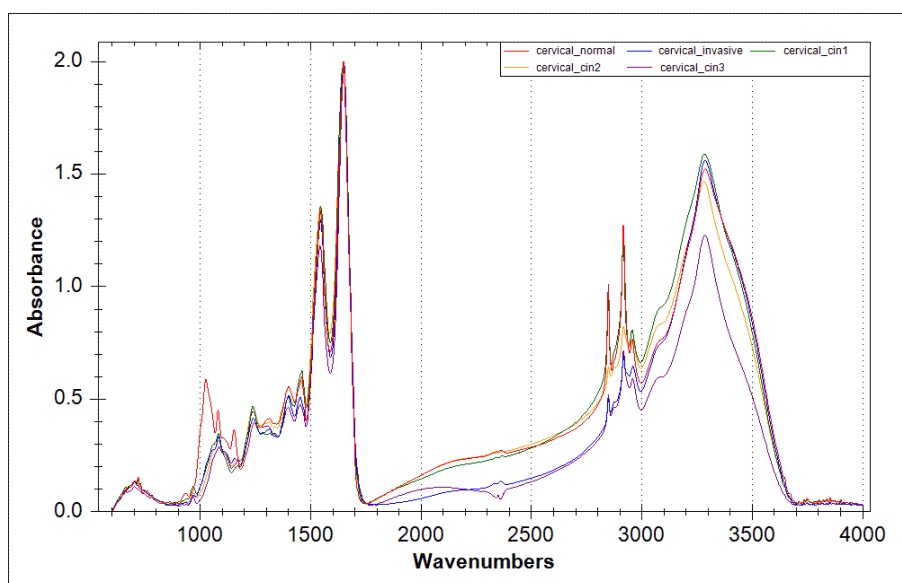


Figure 2. Selected spectra of various stages of cervical neoplasia (normal, CIN1, CIN2, CIN3 and invasive)

The first and simplest step towards demonstrating the differences between stages would be to subtract each spectrum from the known "healthy" section. In the action selection box, we choose 'Subtract', and then following the directions in the Action Description box we select the "normal" spectrum at the top spectrum selection and each of the other stages at the bottom spectrum selection before pressing the 'Perform' button. This produces the result seen in Figure 3. We can begin to see the differences between stages here, mainly between 1700 - 2800 and 3000 - 3600 cm^{-1} wavenumbers, but defining characteristics of each stage are not clear yet.

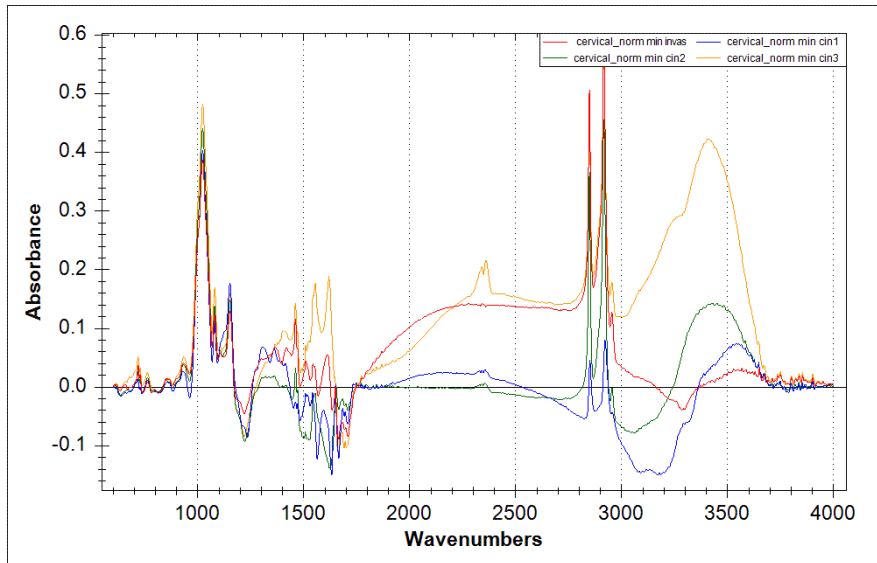


Figure 3. Comparison between the difference spectra (between normal and invasive, CIN1, CIN2, or CIN3)

We next perform a Principle Component Analysis (PCA) on each of the spectra (normal, invasive, CIN1 - 3). As mentioned earlier with the 'Subtract' action, we follow similar steps to select 'PCA' from the action list, and following the Action Description we select each spectrum in the top spectrum selection box, then press 'Perform'. We are then prompted to select from among the eigenvalue/eigenvector pairs calculated by the PCA Figure 4. After selecting both, a new transformed 'spectrum' is generated and added to the plot with the suffix "pca". Our new coordinate axes represent the variation along the chosen eigenvectors, which shows the significance of each original datum. We repeat this with each of the other base spectra to end up with five new transformed spectra. However, while we can still see some differences between them, it is not much more informative than the original spectra. When we subtract each suspicious PCA "spectrum" from the normal, the differences become more clear (Figure 5.), and we can use their positions and alignments to help determine their stage (class). At the low end of the PCA "scale" (negative x-values), we see that advanced stages of CIN are higher on the y-axis (with the exception of the invasive neoplasia), while on the positive side we see that the advanced stages appear lower.

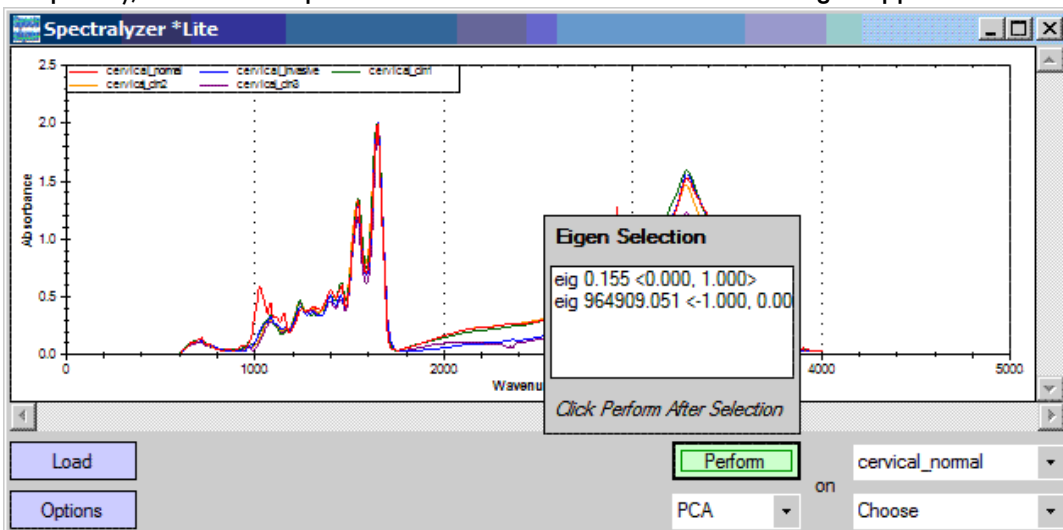


Figure 4. Performing the Principle Component Analysis (PCA) - selecting eigenvalues/eigenvectors

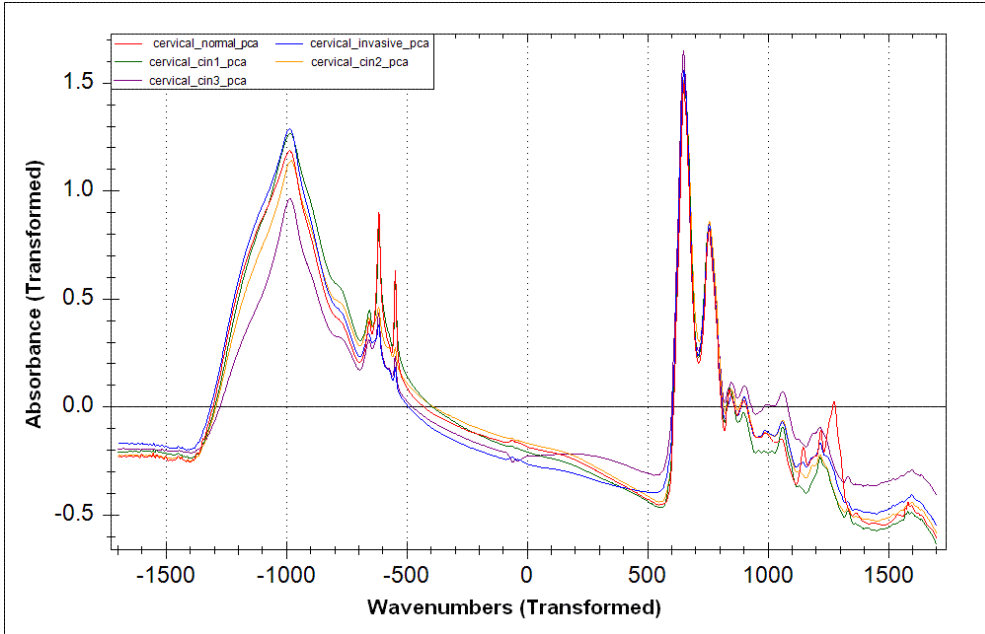


Figure 5. Spectra transformed by the Principle Component Analysis (PCA). We note that the coordinate space has also been transformed.

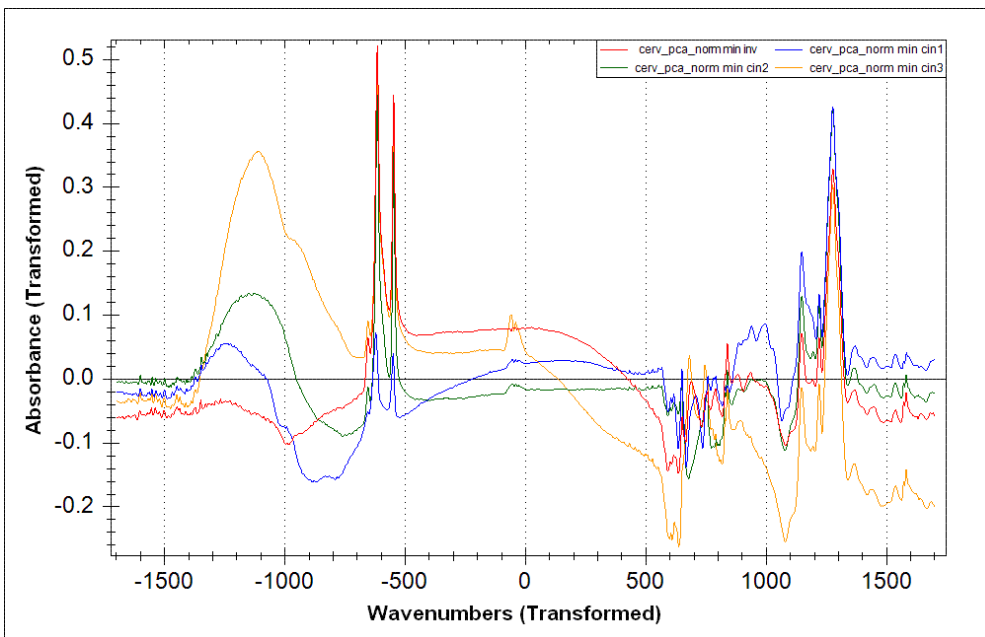


Figure 6. Differences PCA spectra for each stage with respect to normal. The wavenumber axes has been transformed

We can further differentiate the spectra using Linear Discriminant Analysis. Again, as easily as before, we select the action 'LDA' and follow the instructions (Figure 7); performing the analysis on each stage (bottom selection) with respect to the "normal" spectrum (top selection).

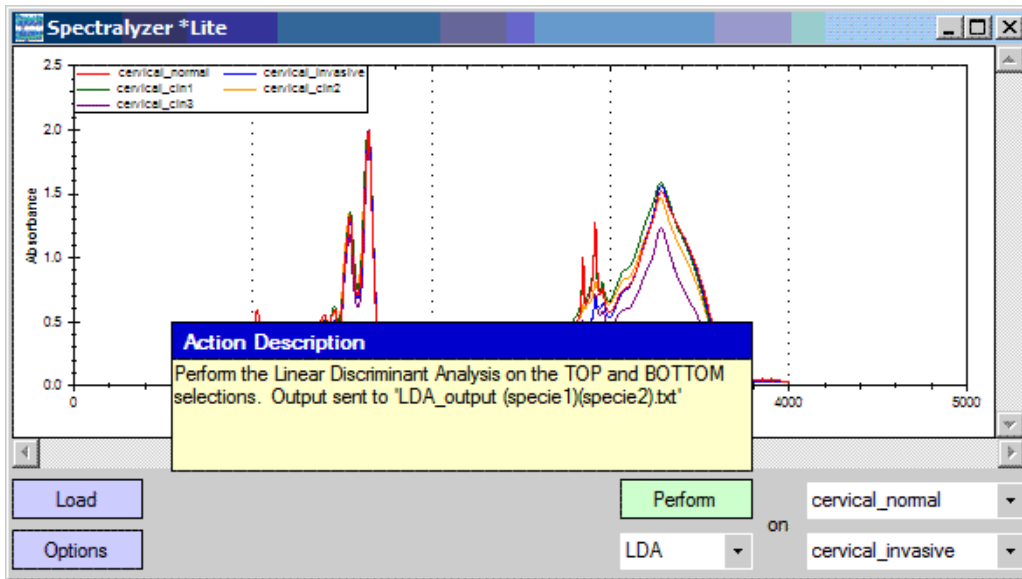


Figure 7. Performing Linear Discriminant Analysis (LDA) -- instructions shown, before saving the analysis file

LDA creates three new spectra for each operation -- the transform of each original spectrum, and the threshold between the two. The new coordinate axes represent how the "object feature" (X) determines the "categorized group" (Y) based on the original axes. The results of the LDA are automatically saved in a text file of the form "LDA output (specie1)(specie2).txt", containing the new coordinates and their "reclassification" as well as a confusion table describing the accuracy of the LDA predictions. When we examine the transformed groupings, we can immediately see a gradual difference in the progression of the neoplasia (Figure 8). From the LDA, there are three distinct areas of the transformed spectra -- a lower region between 0 and -1, the middle region between 0 and 4, and an upper section above 4. As the neoplasia advances from normal to CIN3, the three regions show increasing amplitude around their respective threshold lines, but when it progress to the invasive stage the amplitudes decrease slightly.

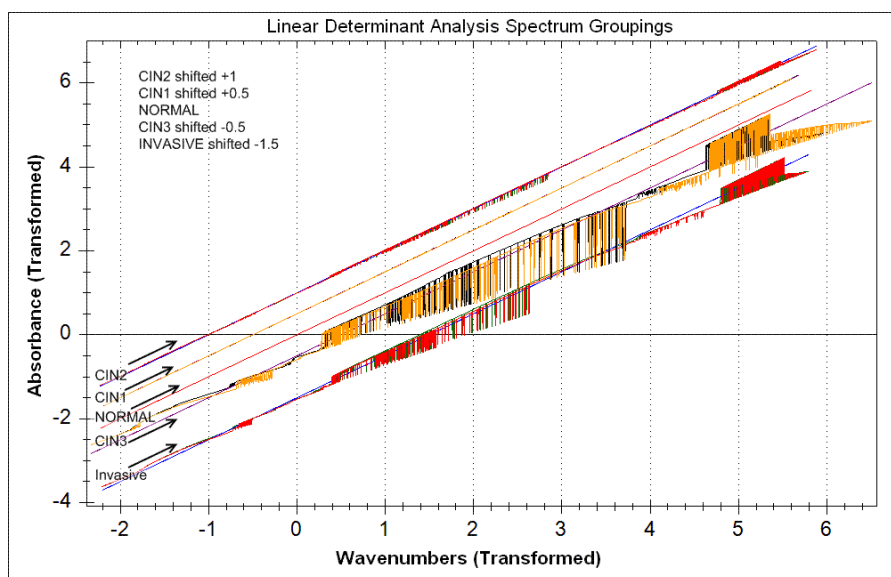


Figure 8. Transformed "spectra" from the LDA; spectra are shifted arbitrarily as indicated in the figure to improve visibility. The wavenumbers axes is transformed

8. Conclusion

The PAP smear, which is currently used for initial screening for cervical cancer, has the major drawback of false negatives, and its analysis is time consuming. Thus other methods that are objective, rapid, and less prone to error are desirable. Changes in biochemical compositions can be an indication of cancer [32,33]. However, such studies are expensive and complicated [34]. Earlier research highlights the importance of detection of biochemical changes in formalin-fixed tissue for diagnosis of cervical cancer through the use of Fourier-transform infrared micro spectroscopy. The changes manifested in the FTIR spectra were used as fingerprints with which to classify and diagnose the various types and stages of disease that are not easily detectable through conventional methods.

Thus approved techniques without an automatic, objective, sensitive and rapid analyze tool are not sufficient, and it is for this reason that the SPECTRALYZER was developed as a complementary. In effect it constitutes a tool that melds traditional analysis methods with advanced computational methods such as the Principal Component and the Linear Discriminant Analyses which are comparable to the gold standard approach.

In Cervical Cancer diagnosis test case, the meld of the Principal Component and the Linear Discriminant Analyses performed by the SPECTRALYZER program exhibit promising potential for an earlier, more accurate diagnosis of cervical neoplasia, and should certainly be investigated further. For other ailments with biochemical changes, accurate microscopic FTIR data in tandem with the SPECTRALYZER program may lead to earlier detection, critical for identifying these conditions before they become untreatable.

9. Downloading and installing SPECTRALYZER

News, updates, and all versions of SPECTRALYZER are all available from our web site : <http://nmcrc.sce.ac.il>.

To install SPECTRALYZER, the compressed file can be downloaded from the NMCRC website. It needs only to be unzipped to a folder of the user's choice, and can be run immediately by running the executable file.

10. References

1. American Cancer Society **Cancer Facts and Figures 2002**, American Cancer Society, Atlanta, Ga., 2002
2. Andrus., P. G. and Strickland, R. D. **Cancer grading by Fourier transform infrared spectroscopy**, *Biospectroscopy*, 4 ,1, 1998, pp. 37-46
3. Argov, S., Ramesh, J., Salman, A., Silenikov, I., Goldstein, J., Guterman, H. and Mordechai, S. **Diagnostic potential of Fourier transformed infrared microspectroscopy and advanced computational methods in colon cancer patients**, *Journal of Biomedical Optics*, 7(1), 2002, pp. 248-254
4. Bauer, H. M. , Ting, Y., Greger, C. E., Chambers, J. C., Tashiro, C. J., Chimera, J., Reingold, A. and Manos, M. M. **Genital human papillomavirus infection in female university students as determined by a PCR-based method**, *J. Am. Med. Assoc.*, **265**, 1991, pp. 472-477
5. Bell, R. J. **Introductory Fourier Transform Spectroscopy**, Academic Press, NY, 1972

6. Booch, G. et al. **UML Users' Guide**, Reading, MA: Addison-Wesley-Longman, 2000
7. Booch, G., Rumbaugh, J. and Jacobson, I. **The Unified Modeling Language User Guide**, Addison Wesley, 1998
8. Chiriboga, L., Xie, P., Yee, H., Zarou, D., Zakim, D. and Diem, M. **Infrared spectroscopy of human tissue. IV. Detection of dysplastic and neoplastic changes of human cervical tissue via infrared microscopy**, Cellular Molecular Biology, Noisy-le-Grand, France, **44**, 1998, pp. 219-229
9. Crum, C. P. **Female genital tract**, Robbins Pathologic basis of disease 5th edition, Cotran, R. S. et al. (eds.), Philadelphia, 1994, pp. 1049-1052
10. Fields, A. B., Jones, J. G., Thomas, G. M. and Runowicz, C. D. **Gynecologic cancer**, Clinical Oncology, Lenhard, R., Osten, E. R. T. and Gansler, T. (eds.), American Cancer Society, Blackwill Science, Inc., 2001, pp. 455-497
11. Fowler, M. **Analysis Patterns**, Addison-Wesley, 1997
12. Gamma, E., Helm, R., Johnson, R. and Vlissides, J. **Design Patterns - Elements of Reusable Object-Oriented Software**, Addison-Wesley, 1994
13. Georgakoudi, I., Jacobson, B. C., Miller, M. G., Sheets, E. E., Badizadegan, K., Carr-Locke, D. L., Crum, C. P., Boone, C. W., Dasari, R. R., Van Dam, J. and Feld, M. S. **NAD(P)H and Collagen as In Vivo Quantitative Fluorescent Biomarkers of Epithelial Pre-Cancerous Changes**, Cancer Research, **62**, 2002, pp. 682-687
14. Ireland, A. **Software Engineering 4: The Software Testing Life-Cycle**, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, 2007
15. Jacobson, I., Christerson, M., Jonsson, P. and Overgaard, G. **Object-Oriented Software Engineering - A Use Case Driven Approach**, Addison Wesley, 1992
16. Kumar, A., Sharma, S., Pundir, C. S. and Sharma, A. **Decreased plasma glutathione in cancer of the uterine cervix**, Cancer Lett., **94**, 1995, pp. 107-111
17. Larman, C. **Applying UML and Patterns**, Prentice Hall, 2nd. Ed., 2000
18. MacDonald, M. **User Interfaces in C#: Windows Forms and Custom Controls**, Apress, US, 2002
19. Manju V, Sailaja JK, Nalini N. **Circulating lipid peroxidation and antioxidant status in cervical cancer patients: a case-control study**, Clin. Biochem., **35**(8), 2002, pp. 621-625
20. Mark, S., Sahu, R. K., Kantarovich, K., Putshibalov, A., Guterman, H., Goldstien, J., Jagannathan, R., Argov, S. and Mordechai, S. **Fourier transform infrared microspectroscopy as a quantitative diagnostic tool for assignment of premalignancy grading in cervical neoplasia**, Journal of Biomedical Optics, **9**(3), 2004, pp. 558-567
21. Meyer, B. **Object-Oriented Software Construction** (2nd. ed.), Prentice Hall, 1997
22. Mordechai, S. et. al. **Fourier Transform Infrared Micro spectroscopy as a Quantitative Diagnostic Tool For Assignment of Premalignancy Grading in Cervical Neoplasia**, Pending
23. Morris, B. J., Lee, C., Nightingale, B. N., Molodysky, E., Morris, L. J., Sternhell, S., Cardona, M., Mackerras, D. and Irwig, L. M. **Fourier transform infrared spectroscopy of dysplastic, papillomavirus-positive cervicovaginal lavage specimens**, Gynecol. Oncol. **56** (2), 1995, pp. 245-249
24. Pudshyvalov, A., Mark, S., Sahu, R. K., Kantarovich, K., Guterman, H., Goldstien, J., Jagannathan, R., Argov, S. and Mordechai, S. **Distinction of cervical cancer biopsies by use of infrared microspectroscopy and probabilistic neural networks**, Applied Optics **44**(18), 2005, pp. 3725-3734
25. Radatz, J., Olson, M. and Campbell, S. **MIL-STD-498**, Logicon, 1995 Feb
26. Ramanujam, N. **Fluorescence spectroscopy in vivo**, In *Encyclopedia of Analytical Chemistry*, Meyers, R. A. (ed.), John Willey & Sons, Ltd., Chichester, New York, 2000, pp. 20-56

27. Ramanujam, N., Mitchell, M. F., Mahadevan, A., Thomsen, S., Silva, E. and Richards-Kortum, R. **Fluorescence Spectroscopy: a Diagnostic Tool for Cervical Intraepithelial Neoplasia (CIN)**, *Gynecologic Oncology*, **52**, 1994, pp. 31-38
28. Rumbough, J., Blaha, M., Permerlani, W., Eddy, F. and Lorensen, W. **Object-Oriented Modeling and Design**, Prentice Hall, 1991
29. Schiffman, M. H., Brinton, L. A., Devesa, S. S., Fraumeni, J. and Joseph, F. **Cervical cancer**, in "Cancer Epidemiology and Prevention", Schottenfeld, D., Fraumeni, J. and F. Joseph, F. (eds.), Oxford U. Press, New York, 1996
30. Sherrod, P. **DTREG: Software For Predictive Modeling and Forecasting**, 2003, online source: <http://www.dtreg.com/DTREG.pdf>
31. Smith, L. I. **A tutorial on Principal Components Analysis**, 26 Feb 2002, online source: <http://moodle.epfl.ch/mod/resource/view.php?id=30011>
32. Teknomo, K. **Discriminant Analysis Tutorial**, 2006, online source: <http://people.revoledu.com/kardi/tutorial/LDA/index.html>
33. Watkins, J. **Testing IT: An Off-the-Shelf Software Testing Process**, Cambridge University Press, 2001 May 1
34. Wirfs-Brock, R., Wilkerson, B. and Wiener, L. **Designing Object-Oriented Software**, Prentice Hall, 1990

¹**Jeremy Schwartz** is a software developer with a broad range of programming experience from research, consulting, and academic settings. He graduated in 2006 from North Carolina State University (USA) with undergraduate degrees in Chemical Engineering and Computer Science. Mr. Schwartz specializes in writing highly functional, well-documented and efficient code for both online and offline environments. Having written training materials, developed data analysis tools, and rebuilt several websites from the ground up, he understands the iterative process by which user requirements are transformed into deliverables.

²**Shlomo Mark** is a senior lecturer at SCE Shamon college of engineering. He earned his Ph.D. in nuclear engineering and an M.Sc. in **Biomedical Engineering and in Managing and Safety Engineering**. He works in the Department of Software Engineering. and he is the Head of the NMCRC - Negev Monte Carlo Research Center, Shamon College of Engineering. His main research interests are Scientific programming, Computational modeling for Physical, environmental and medical applications, Monte Carlo Simulations, Develop, upgrade, and improved Monte Carlo based codes, by using software engineering mythologies.

³ Corresponding author. Tel.: +972 8 647 5631; fax: +972 8 647 5623

⁴ Codification of references:

[1]	American Cancer Society Cancer Facts and Figures 2002 , American Cancer Society, Atlanta, Ga., 2002
[2]	Andrus., P. G. and Strickland, R. D. Cancer grading by Fourier transform infrared spectroscopy , <i>Biospectroscopy</i> , 4 , 1, 1998, pp. 37-46
[3]	Argov, S., Ramesh, J., Salman, A., Silenikov, I., Goldstein, J., Guterman, H. and Mordechai, S. Diagnostic potential of Fourier transformed infrared microspectroscopy and advanced computational methods in colon cancer patients , <i>Journal of Biomedical Optics</i> , 7 (1), 2002, pp. 248-254
[4]	Bauer, H. M., Ting, Y., Greer, C. E., Chambers, J. C., Tashiro, C. J., Chimera, J., Reingold, A. and Manos, M. M. Genital human papillomavirus infection in female university students as determined by a PCR-based method , <i>J. Am. Med. Assoc.</i> , 265 , 1991, pp. 472-477
[5]	Bell, R. J. Introductory Fourier Transform Spectroscopy , Academic Press, NY, 1972
[6]	Booch, G. et al. UML Users' Guide , Reading, MA: Addison-Wesley-Longman, 2000
[7]	Booch, G., Rumbaugh, J. and Jacobson, I. The Unified Modeling Language User Guide , Addison Wesley, 1998
[8]	Chiriboga, L., Xie, P., Yee, H., Zarou, D., Zakim, D. and Diem, M. Infrared spectroscopy of human tissue. IV. Detection of dysplastic and neoplastic changes of human cervical tissue via infrared microscopy , <i>Cellular Molecular Biology</i> , Noisy-le-Grand, France, 44 , 1998, pp. 219-229
[9]	Crum, C. P. Female genital tract , Robbins Pathologic basis of disease 5 th edition, Cotran, R. S. et al. (eds.), Philadelphia, 1994, pp. 1049-1052
[10]	Fields, A. B., Jones, J. G., Thomas, G. M. and Runowicz, C. D. Gynecologic cancer , Clinical Oncology, Lenhard, R., Osten, E. R. T. and Gansler, T. (eds.), American Cancer Society, Blackwill Science, Inc., 2001, pp. 455-497
[11]	Fowler, M. Analysis Patterns , Addison-Wesley, 1997
[12]	Gamma, E., Helm, R., Johnson, R. and Vlissides, J. Design Patterns - Elements of Reusable Object-Oriented Software , Addison-Wesley, 1994

[13]	Georgakoudi, I., Jacobson, B. C., Miller, M. G., Sheets, E. E., Badizadegan, K., Carr-Locke, D. L., Crum, C. P., Boone, C. W., Dasari, R. R., Van Dam, J. and Feld, M. S. NAD(P)H and Collagen as In Vivo Quantitative Fluorescent Biomarkers of Epithelial Pre-Cancerous Changes , <i>Cancer Research</i> , 62 , 2002, pp. 682-687
[14]	Ireland, A. Software Engineering 4: The Software Testing Life-Cycle , School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, 2007
[15]	Jacobson, I., Christerson, M., Jonsson, P. and Overgaard, G. Object-Oriented Software Engineering - A Use Case Driven Approach , Addison Wesley, 1992
[16]	Kumar, A., Sharma, S., Pundir, C. S. and Sharma, A. Decreased plasma glutathione in cancer of the uterine cervix , <i>Cancer Lett.</i> , 94 , 1995, pp. 107-111
[17]	Larman, C. Applying UML and Patterns , Prentice Hall, 2nd. Ed., 2000
[18]	MacDonald, M. User Interfaces in C#: Windows Forms and Custom Controls , Apress, US, 2002
[19]	Manju V, Sailaja JK, Nalini N. Circulating lipid peroxidation and antioxidant status in cervical cancer patients: a case-control study , <i>Clin. Biochem.</i> , 35 (8), 2002, pp. 621-625
[20]	Mark, S., Sahu, R. K., Kantarovich, K., Putshibalov, A., Guterman, H., Goldstien, J., Jagannathan, R., Argov, S. and Mordechai, S. Fourier transform infrared microspectroscopy as a quantitative diagnostic tool for assignment of premalignancy grading in cervical neoplasia , <i>Journal of Biomedical Optics</i> , 9 (3), 2004, pp. 558-567
[21]	Meyer, B. Object-Oriented Software Construction (2nd. ed.), Prentice Hall, 1997
[22]	Mordechai, S. et. al. Fourier Transform Infrared Micro spectroscopy as a Quantitative Diagnostic Tool For Assignment of Premalignancy Grading in Cervical Neoplasia , Pending
[23]	Morris, B. J., Lee, C., Nightingale, B. N., Molodysky, E., Morris, L. J., Sternhell, S., Cardona, M., Mackerras, D. and Irwig, L. M. Fourier transform infrared spectroscopy of dysplastic, papillomavirus-positive cervicovaginal lavage specimens , <i>Gynecol. Oncol.</i> 56 (2), 1995, pp. 245-249
[24]	Pudshyvalov, A., Mark, S., Sahu, R. K., Kantarovich, K., Guterman, H., Goldstien, J., Jagannathan, R., Argov, S. and Mordechai, S. Distinction of cervical cancer biopsies by use of infrared microspectroscopy and probabilistic neural networks , <i>Applied Optics</i> 44 (18), 2005, pp. 3725-3734
[25]	Radatz, J., Olson, M. and Campbell, S. MIL-STD-498 , Logicon, 1995 Feb
[26]	Ramanujam, N. Fluorescence spectroscopy in vivo , In <i>Encyclopedia of Analytical Chemistry</i> , Meyers, R. A. (ed.), John Willey & Sons, Ltd., Chichester, New York, 2000, pp. 20-56
[27]	Ramanujam, N., Mitchell, M. F., Mahadevan, A., Thomsen, S., Silva, E. and Richards-Kortum, R. Fluorescence Spectroscopy: a Diagnostic Tool for Cervical Intraepithelial Neoplasia (CIN) , <i>Gynecologic Oncology</i> , 52 , 1994, pp. 31-38
[28]	Rumbough, J., Blaha, M., Permerlani, W., Eddy, F. and Lorensen, W. Object-Oriented Modeling and Design , Prentice Hall, 1991
[29]	Schiffman, M. H., Brinton, L. A., Devessa, S. S., Fraumeni, J. and Joseph, F. Cervical cancer , in "Cancer Epidemiology and Prevention", Schottenfeld, D., Fraumeni, J. and F. Joseph, F. (eds.), Oxford U. Press, New York, 1996
[30]	Sherrod, P. DTREG: Software For Predictive Modeling and Forecasting , 2003, online source: http://www.dtreg.com/DTREG.pdf
[31]	Smith, L. I. A tutorial on Principal Components Analysis , 26 Feb 2002, online source: http://moodle.epfl.ch/mod/resource/view.php?id=30011
[32]	Teknomo, K. Discriminant Analysis Tutorial , 2006, online source: http://people.revoledu.com/kardi/tutorial/LDA/index.html
[33]	Watkins, J. Testing IT: An Off-the-Shelf Software Testing Process , Cambridge University Press, 2001 May 1
[34]	Wirfs-Brock, R., Wilkerson, B. and Wiener, L. Designing Object-Oriented Software , Prentice Hall, 1990

THE PROBABILITY MODEL FOR RISK OF VULNERABILITY TO STDs/OR HIV INFECTION AMONG PRE-MARITAL FEMALE MIGRANTS IN URBAN INDIA

Himanshu PANDEY

Department of Mathematics and Statistics,
D.D.U. Gorakhpur University, Gorakhpur (U.P.), India

E-mail: himanshu_pandey62@yahoo.com

Kamlesh Kumar SHUKLA

Institute of Management Studies, Lal Quan,
Bulandshahar Road, Ghaziabad-201009, National Capital Region, India

E-mail:

Abstract: *In the study, authors have proposed a mathematical model for unmarried female migrant workers having number of closed boy friends. They are more vulnerable to STDs and HIV transmission. The model is fitted well on the given data and estimate of female migrants having one close boy friend was found maximum. The study based on 362 pre-marital female migrant workers less than 30 years of age in Delhi urban India, while they have wanted lavish life styles and having number rich boy friends.*

Key words: *STDs; Vulnerability to HIV; closed boy friend*

1. Introduction

India, it is well known that females have been participating equally as males now days. , women have started taking up professional roles and they are now entering new fields such as administration, science, technology, medicine, journalism and etc. We all know that India is a male dominated country, but we should remove this word from our dictionary. Historically, females are totally dependent on the males. Now, we see that females have been migrating on their own irrespective of their martial status. The question arises, why do we need such study? Answer is simple that STDs and HIV/AIDS are increasing. It is not a particular answer of this question. It is major problem of world and not of India. HIV/AIDS is not clearly related to female migrants, it is related to male migrants also, but some how related to migrants. (Reddy, 2004) A study of 120 samples of HIV/AIDS infected people was taken from Sur Sunderlal Hospital, Banaras Hindu University, India; the study reported that about 70% of them were migrants. Now (Jain, et al 2007) National AIDS Control Organization (NACO) estimated around 0.6 million HIV infections in 2002. They also estimated that between 3.8 and 4.5 million Indians were living with HIV/AIDS during 2002, of who around 39 percent were women. The epidemic continues to shift towards women and young people.

According to an estimate of UNAIDS, although HIV prevalence rate is low (around 1 percent), the overall number of people with HIV infection is high. The majority of the reported AIDS cases have occurred in the sexually active and economically productive age group. Earlier men were the main transmitters of the disease but now studies are showing that females are also transmitting the disease to males. A study conducted among 379 HIV-infected people in 1991, reported in the journal of the American Medical Association, observes an evidence of female-to-male transmission.

Women are working in almost all types of jobs, such as technical, professional and non-professional in both private and public sectors. So, the traditional role of women as housewives has gradually changed into working women and housewives (Reddy, 1986; Anand, 2003).

The real world phenomenon indicates three distinct types of female migration (Fawcett et al, 1984) (a) Autonomous female migration: Many middle and upper middle class women migrate to cities for improving their educational credentials and also to get suitable employment apparently in a quest for social advancement and also to enhance their status in the marriage market. Among the semi-literate, young girls migrating to towns/cities to work in export processing units, garment industry, electronic assembling and food processing units is continuously on the increase in the recent years; (b) Relay migration: To augment family income, families which have some land holdings in the rural area, send the daughters to work mostly as domestic servants where they are safe in the custody of a mistress. First the elder daughter is sent out and she is replaced by the second, third and so on, as one by one get married.; (C) Family migration: Here the wife instead of staying back in the village prefers to join her husband in the hope of getting some employment in the destination area. Family migration among agricultural wage labourers who have no land or other assets to fall back at times of crisis is becoming increasingly common. Qian,X;et al.(2005) have also been studied in a large population of young people (age 10-24 years) in the region of Asia and the Pacific. Adjusting to sexual development and protecting their reproductive health are among the greatest challenges for adolescents during this period of transition from childhood to adulthood.

When women get empowered, they benefit themselves and the larger community (Hugo 2000). 'The expansion of women's A capability not only enhances women's own freedom and well-being but also has many other effects on the lives of all. An enhancement of women's active agency can in many circumstances contribute substantially to the lives of all people –men, women and children as well as adults' (Sen, 2001). Hongjie Liu,et al, (2005)have reported in the study, which is as under : the study was to identify risk factors associated with sexually transmitted diseases (STDs) among rural-to-urban migrants in Beijing in 2002. Migrants with STDs consisted of 432 migrants who sought STD care in two public STD clinics. Migrants without STDs included 892 migrants recruited from 10 occupational clusters. Compared to migrants without STDs, migrants with STDs were more likely to report having engaged in commercial sex (selling or buying sex) (odds ratio [OR] _ 2.70, 95% confidence interval [CI]: 1.71–4.25), multiple sex partners in the previous month (OR _ 6.50, 95% CI: 3.73–11.32) and higher perceived HIV-related stigma (OR _ 1.89, 95% CI: 1.30–2.75). Being a migrant with an STD was also associated with female gender (OR _ 4.10, 95% CI: 2.89–5.82), higher education (OR _ 2.92, 95% CI: 1.40–6.06), and higher monthly salary (OR _ 1.68. 95% CI: 1.23–2.29). Migrants with STDs visited their hometowns more frequently and had more stable jobs than migrants without STDs. Approximately 10%

of the migrants with STDs and 7.7% of the migrants without STDs always used condoms. This indicates that among migrants, acquisition of an STD is associated with higher participation in risk behaviors as would be expected, but also with higher perceived stigma, education, stable jobs, salary, and with female gender.

In the Indian context women in the migrant households do play an important role in family survival but unfortunately they remain invisible in the official data because of the way the concepts are defined and data is collected. But the limited research studies that are available in this concern for the earlier periods indicate that these women are exposed more to the risk of sexual harassment and exploitation (Acharya, 1987 and Saradmoni, 1995). Women migrant workers in sugarcane cutting, work almost twenty hours a day (Teerink, 1995) Female labour mostly from Kerala in the fish processing industries in Gujarat are subject to various forms of hardship and exploitation at the hands of their superiors (Saradmoni, 1995).

Among females, the proportion of migrant and non-migrant workers in white-collar jobs was almost similar in 1971 but the same became smaller in 1991 than that of the non-migrant workers. There are more migrant women than the non-migrant women in the category of blue-collar jobs (Premi, 2001). A study conducted by Dholakia and Dholakia (1971) for 20 major Indian states showed that per capita income, average size of households and overall literacy rates were the main factors explaining the variations in female participation rates across the states. Hirsch, Jennifer S. et al. (2002) have collected data on involved life histories and participant observation with migrant women in Atlanta and their sisters or sisters-in-law in Mexico and the reported that both younger and older women acknowledged that migrant men's sexual behavior may expose them to HIV and other sexually transmitted diseases. Younger Mexican women in both communities expressed a marital ideal characterized by mutual intimacy, communication, joint decision making, and sexual pleasure, but not by willingness to use condoms as an HIV prevention strategy.

Involvement, in risky behaviour can have negative repercussions on their health. In the case of unmarried women after marriage this burden of disease may be transmitted to their husbands and children as well (Jain, et al ,2006). Finally, the growing evidence of an association between migration and risky behaviour (UNAIDS and IOM, 1998), as well as the entry of sexually active migrant working women into the urban areas each year (Visaria, 1998), point to a need for a new sense of urgency.

Different types of mathematical models have been used by different statistician to represents the observed phenomenon in a concise form and systematic approach for different group of migrant's households.

A good number of studies took place after models have been proposed to study the pattern of rural out migration of male greater than 15 years of age and total number of migrants (including wives and children) out migration (Iwunor, 1995, Singh, Yadav, 1991 and Shukla and Yadav, 2006).

In the present paper authors have proposed probability model based on the above studies and applied to risk for sexually transmitted infections or HIV transmission or unwanted pregnancies due to change in sexual behaviour of single female migrants. They have wanted to make different closed boy friends and most of them have been taking interest to go out with friends for movies or drama or to restaurants or hotels, while some of them go to night clubs, discos, bars, pubs or attend late night parties.

2. Data

The study is based on surveyed of 362 unmarried working women, randomly selected from 12 working women's hostels in Delhi. The list of the hostels was obtained from Social Welfare Department, YWCA and wardens of the hostels. Details about the data are given in Jain, et al, 2007.

3. Model

The present model is based on displaced geometric distribution. Proposed model for the number of closed boy friends to describe the distribution of single unmarried female migrants.

- (i). Let α be the proportion of female migrants having at least one closed boy friend.
- (ii) Out of α proportion of female migrants, let β be the proportion of female migrants having only one closed boy friends.
- (iii) Number of closed boy friends attached with female migrants follows a displaced Geometric distribution.
- (iv) Let p be the probability of closed boy friends attached with young unmarried female migrants, they are more vulnerable to STDs/HIVs infections.

Under the above assumptions, the probability distribution for number of closed boy friends, x (say) in given by

$$P(x=k) = \begin{cases} 1 - \alpha & \text{if } k = 0 \\ \alpha\beta & \text{if } k = 1 \\ (1 - \beta)\alpha pq^{k-2} & \text{if } k = 2 \end{cases} \quad (1)$$

The above model involves three parameters, α , β and p to be estimated from observed distribution of female migrants.

These are estimated by equating theoretical frequencies to the observed frequencies of first and second cells and theoretical mean to the observed mean

$$\begin{aligned} \text{i.e.} \quad 1 - \alpha &= \frac{N_0}{N} \\ \alpha \beta &= \frac{N_1}{N} \text{ and} \\ \alpha \beta + (1 - \beta) \alpha ((1-p)/p) &= \text{mean } (x) \end{aligned}$$

4. Result and discussion

The table -A shows that the distribution of observed and expected number of young female migrants and their number of closed boy friends. The value of $\chi^2 = 7.67$ was found insignificant at 1% level. This indicates that proposed model fitted well to the distribution of female migrants. An estimate of the proportion of female migrants having only one closed

boy friend was found very low (0.4364) in comparison to having at least one closed boy friend (0.8039). While, most of the young female migrants believe that their friends watch pornographic material such as blue films, sexy material on the internet, pornographic magazines, posters, photos, etc. and some of them agreed that they had exposure to pornographic material and more often on their computer or DVD player with their friends. Fewer knew that condom use can prevent STDs and HIV/ AIDS. Majority of respondents (86%) did not feel that a healthy looking person could have AIDs (Jain, et al 2007). This study indicates that number female (0.8039) migrants are having close boy friends. They are more vulnerable to STDs and HIV/AIDS. The first sexual event has clear health implications, since it marks initiation into the sexual act which if unprotected, and carries a risk of adverse outcomes such as unplanned pregnancy, HIV and sexually transmitted infections (Wellings et al. 1994).

Table 1. Observed and Expected numbers of unmarried single female migrants according to their close boy friends

No. of closed boy friends	Observed	Expected
0	71	71
1	127	127
2	60	93.4
3	55	40.5
4	19	17.7
5+	10	12.4
Total	362	362

$$\chi^2 = 7.67$$

$$\alpha = 0.8039$$

$$\beta = 0.4364$$

$$p = 0.5691$$

$$d.f. = 2$$

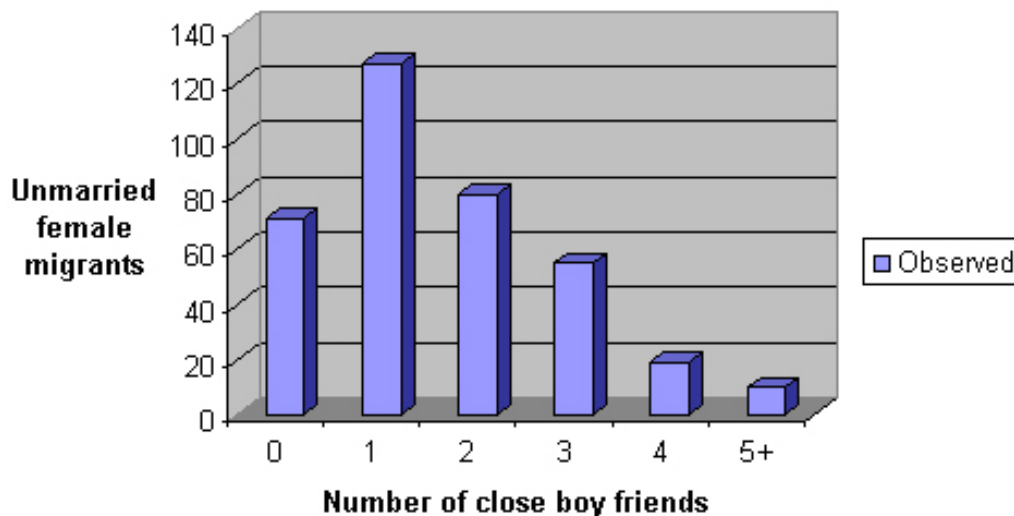


Figure 1. Number of close boy friends and associated with unmarried female migrants

5. Conclusions

A model is proposed and tested a set of data on the number young unmarried female migrant workers less than 30 years of age having number of closed boy friends. They are closely attached with their boy friends. Proposed model is fitted satisfactorily. An estimate of female migrants having at least one boy friend has been found maximum, it is two times than female migrants having one closed boy friends, Hence increasing the living and working condition and they need adequate support should be provided to single young migrants women that might be make them vulnerable to STDs and HIV/AIDS infection.

References

1. Anand, N. **Working Women: Issues and Problems**, *Yojana*, 47, 3, 2003, pp. 11-14
2. Binoy, A. **Beyond Bricks and Stone**, PRIA, New Delhi, 1987
3. Breman, J. **Of Peasants Migrants and Paupers: Rural Labour Circulation and Capitalist Production in Western India**, Oxford University Press, New Delhi, 1985
4. Brown, T. **The Asian Epidemic model: a process model for exploring policy and programme alternatives in Asia**, *Sex Transm. Infect.*, Vol. 80, 2004, pp. 19-26
5. Dholakia, B. H. and Dholakia, R. H. **Interstate Variations in Female Labour Force Participation Rates in India**, *Journal of Labour Economic*, 1978
6. Fawcett, J. T., Khoo, S. and Smith, P. C. **Women in the Cities of Asia: Migration and Urban Adaptation**, Westview Press Boulder Colorado, 1984
7. Hirsch, J. S. et al. **The Social Constructions of Sexuality: Marital Infidelity and Sexually Transmitted Disease-HIV Risk in a Mexican Migrant Community**, *American Journal of Public Health*, Vol. 92, No. 8, 2002, pp. 1227-1237
8. Hongjie, L. et al. **Risk Factors for Sexually Transmitted Disease among rural to urban migrants in China: Implications for HIV/ Sexually Transmitted Disease Prevention**, *Aids Patient Care and STDs*, Vol. 19, No.1, 2005
9. Hugo, G. **Migration and Women's Empowerment**, in H. B. Presser and G. Sen (eds.), *Women's Empowerment and Demographic Processes*, Oxford University Press, 2000
10. Iwunor, C. C. O. **Estimating of parameter of the inflated geometric distribution for rural out migration**, *Genus*, Vol. L1(3-4), 1995, pp. 253-260
11. Jain, R., Gupta, K. and Singh, A. K. **Sexual Risk Behaviour and vulnerability to HIV infection among young Migrant Women Workers in Urban India**, accessed on July 17, 2007, online source: <http://paa2007.princeton.edu/download.aspx?submissionID=72114>
12. Mahendra, P. K. **Who Migrates to Delhi**, *Demography India*, 30 (1), 2001, pp. 49-59
13. Qian, X., Tan, H., Cheng, H. and Liang, H. **Sexual and reproductive health of adolescents and youths in China: a review of literature and projects from 1995-2002**, World Health Organization Western Pacific Region, 2005
14. Reddy, C. R. **Changing Status of Educated Working Women-A Case Study**, B. R. Publishing Corporation, Delhi, 1986
15. Reddy, D. C. S. **Prevention of HIV/AIDS in Uttar Pradesh**, organized by the state Innovations in Family Planning Services Agency (SIFPSA), U.P. State AIDS control Society and the Policy Project, The Futures Group International, Agra, January 29-31, 2004
16. Rensje, T. **Migration and its impact on Khandeshi Women in the Sugarcane Harvest**, in Schenk Sandbergen, S. (ed.) "Women and Seasonal Labour Migration", IDPAD Sage New Delhi, 1995

17. Saradmoni, K. **Crisis in the Fishing Industry and Women's Migration: The Case of Kerala**, in Schenk Sandbergen (ed.), *Women and Seasonal Labour Migration*, IDPAD Sage, New Delhi, 1995
18. Sen, A. **Many faces of Gender Inequality**, *Frontline*, Vol. 18, No. 22, Oct/Nov, 2001
19. Shukla, K. K. and Yadava, K. N. S. **The distribution of number of migrants at the household level**, *Journal of population and social studies*, Vol.14, (2), Jan 2006
20. UNAIDS and IOM **Migration and AIDS**, *International Migration* 36, 4, 1998, pp. 445-466
21. Visaria, P. **Urbanization in India: Retrospect and Prospect**, Unpublished typescript, 1998
22. Wellings, K., Field, J., Johnson, A. and Wardsworth, J. **Sexual behaviour and Lifestyle in Britain: The national survey of sexual attitudes and lifestyles**, Penguin Books Ltd, London, England, 1994

SUBJECT-LEVEL TREND ANALYSIS IN CLINICAL TRIALS

Alexandru-Ionut PETRISOR

PhD, Assistant Professor, Sections Urbanism and Landscape, School of Urbanism, "Ion Mincu" University of Architecture and Urbanism, Bucharest, Romania

E-mail: alexandru_petrisor@yahoo.com

Web-page: http://www.geocities.com/a_petrisor



Liviu DRAGOMIRESCU

PhD, Associate Professor, Department of Ecology, Faculty of Biology, University of Bucharest, Romania

E-mail: liviu_dragomirescu@yahoo.com



Cristian PANAITI

PhD Candidate, Intern, Diabetes Nutrition and Metabolic Diseases, N. Paulescu Institute, Bucharest, Romania Romania

E-mail: cristian.panaite@gmail.com



Alexandru SCAFA-UDRISTE

PhD Candidate, University Assistant, „Carol Davila” University of Medicine and Pharmacy, Department of Internal Medicine and Cardiology, Clinic Urgency Hospital, Bucharest, Romania

E-mail:



Anamaria BURG

PhD, Assistant Professor, University of South Carolina, Lancaster, S.C., USA

E-mail: a_rusu@hotmail.com



Abstract: *In particular situations, clinical trials researchers could have a potential interest in assessing trends at the level of individual subjects. This paper establishes a common approach and applies it in two different situations, one from nutritional medicine and one from cardiovascular medicine. The approach consists of running as many regression models as the number of subjects, looking at the behavior of some parameter of interest in time. The regression parameters, particularly the slope of the regression line, offer the general sense of the trend and allow for testing its statistical significance. Extrapolation at the level of the entire sample is possible using some version of the binomial test. In both cases, significant results were obtained despite of small sample sizes.*

Key words: *simple linear regression; slope; coefficient of determination; binomial test*

1. Introduction

Clinical trials are used to evaluate new drugs or treatments, including new technologies, assess new screening programs, or ways to organize and deliver health services [1]¹. Epidemiologists distinguish three types of clinical trials: prophylactic - used to prevent diseases, therapeutic - used to treat diseases, and interventional - used to intervene before the disease is developed [2],[3]. Whilst most often clinical trials, similar to other epidemiologic studies, analyze the impact of intervention on the development of a certain disease, assessing the impact of potential risk factors, researchers might be occasionally interested to analyze trends at the level of the subject.

Due to the fact that there are only several important moments when the parameters of interest are checked (beginning of the study, midterm, endpoint and eventually some other intermediary moments), there are too few data to use a time series model. Nevertheless, the evolution of the variable of interest (weight in the first case, in the second) in time translates into a simple linear regression model.

For each regression model, several parameters of interest describe the trend. The sign of the regression slope, β , indicates either a decreasing or increasing trend. Whereas a test of significance for β could pinpoint significant trends in some patients, the limited number of time milestones results into a general lack of significance. The very few significant trends cannot allow for further analyses. The same behavior characterizes the coefficient of determination, R^2 .

To assess the overall trend, a sign test could be used to assess whether the trend is decreasing (a percentage significantly larger than 50% or some other value of the regression slopes are negative) or positive. If sufficient data are available, the values of the regression slopes could be used in conjunction with their significance test and trends can be classified as either significantly decreasing, not significant, or significantly increasing.

Two examples of applying the proposed approach are presented in this paper. In the first case, while comparing the efficiency of three weight loss programs, a question of interest is whether weight loss is consistent during the period when the treatment is administered. Weight is checked at some intervals, and for each patient the efficiency should translate into a continuous decrease of weight. In the second example, two echocardiographic parameters (the ejection fraction and the kinetic score) were analyzed in relationship in a clinical study of the Acute Myocardial Reperfusion Syndrome looking at risk factors, predictors and criteria assessing the success of interventions.

2. Methods

2.1. Statistical tests

The steps taken in applying the proposed methodology are:

1. For each subject, run simple linear regressions according to the model $Y = \alpha + \beta \times \text{TIME}$, where Y is the dependent variable monitored in the study.
2. For each model, record the slope of the regression line, β , or the coefficient of determination, R^2 . Also, test for their significance [6]:

$$t_0 = \frac{\beta}{\sqrt{\frac{\sum_{i=1}^n (Y_i - (\alpha + \beta \cdot X_i))^2}{n-2}} \sqrt{\frac{\sum_{i=1}^n X_i^2 - n \cdot \left(\frac{\sum_{i=1}^n X_i}{n}\right)^2}{n-2}}} \approx t_{n-2} \quad , \text{ or } \quad t_0 = R \sqrt{\frac{n-2}{1-R^2}} \approx t_{n-2} .$$

3. Define an indicator variable to describe the trend as either:
 $I = -1$, if $\beta < 0$; 1 , if $\beta > 0$; and no value otherwise, if β was found significant in very few cases, or
 $I = -1$, if $\beta < 0$ and $p \leq 0.05$; 1 , if $\beta > 0$ and $p \leq 0.05$; and 0 otherwise, if the number of subjects with a significant trend is large enough to allow for further statistical testing.
4. To analyze the overall trend, run the binomial test to compare the proportions of subjects with $I = 1$ and, respectively, -1 (derived from Piegorsch and Bailer [7]):

$$z = \frac{p - 0.5}{0.5/\sqrt{n}} \approx N(0,1)$$

p is the proportion of subjects with $I = 1$ or -1 . 0.5 is the proportion corresponding to the null hypothesis H_0 : there is no overall trend (the proportions of subjects with $I = 1$ and, respectively, -1 are equal, and each of them is 0.5).

5. The indicator variable can be used in the Analysis of Co-Variance or logistic regression, either as dependent or independent variable, depending upon the interest of the researcher.

2.2. Software implementation

The steps described above were implemented in SAS. In order to use SAS, data were stored in an array with the following columns: TIME, the time when each observation was recorded; CLASS, any classification variable (identifying the group to which subjects belong, if any), and S01, S02, ..., S0n, an identifier of each subject (could be automatically generated in Excel). Given this structure, the SAS code is provided below (comments are inserted between accolades `{}`).

```
data name_of_dataset;
input time class S01 S02 ... S0n;
{Paste actual data here}
;
proc reg;
model S01 = time / influence; {The "influence" option was used in the second example to see if
observations recorded at some particular time are more relevant for diagnosis; if there is no interest in
testing it, do not use the "/ influence" statement}
by class; {Should analyses be run only for a particular class, use "where class = class_level" instead; if
there are no classes, do not use any statement}
{Repeat the statements starting with "proc reg;" for S02, ..., S0n}
run;
```

From the SAS output, retrieve for each regression model the values of β and corresponding p-values for the test of $H_0: \beta = 0$ vs. $H_A: \beta \neq 0$. Store all these in Excel in three columns and define the following based on Excel functions, replacing SIGN and BETA with corresponding column names:

The sign of β : SIGN = IF(BETA < 0, "MINUS", IF(BETA = 0, "ZERO", "PLUS"))

The significance of β : SIGNIFICANCE = IF(p < 0.05, "S", "NS")

To compute the proportion of subjects with $\beta > 0$ (respectively $\beta < 0$), use:
 $X = \text{COUNTIF}(\text{RANGE}, \text{"PLUS"})$, respectively $X = \text{COUNTIF}(\text{RANGE}, \text{"MINUS"})$, where X is the position of the cell where the result of the formula is computed and RANGE corresponds to FIRST CELL:LAST CELL of the column where the sign of β is stored.

The sign test can also be computed in Excel using:

$Y = (X/N - 0.5) / (0.5/\sqrt{N})$, where N is the total number of observations.

3. Results and Discussion

3.1. Example #1: Weight loss

Data had been produced by the study "Efficiency of the intensive nutritional, pharmacologic and behavioral management of obesity – correlation of genetic, bio-morphological and psychological factors" (National University Research Council grant #163 of 2006). The aim was to compare the efficiency of three weight loss programs among 84 subjects: classical intervention (24 subjects), intensive intervention assisted by nutritionists (33 subjects), and intensive intervention assisted by psychologists (27 subjects). The later two categories were joined in a group labeled "intensive interventions" (60 subjects).

In this case, the variable of interest was actual weight, recorded in the beginning of the study (0 months), during the study, after 1 and respectively 4 months, and in the end of the study (12 months). Its values were not recorded uniformly, and the actual sample sizes were diminished (Table 1).

Table 1. Subject level linear regression coefficients of weight variation in time: Bucharest, 2008

Classical intervention			Intensive intervention assisted by nutritionists			Intensive intervention assisted by psychologists		
β	$t(\beta)$	$p(t)$	β	$t(\beta)$	$p(t)$	β	$t(\beta)$	$p(t)$
-0.52	5.74	0.1099	-0.54	0.85	0.5532	-0.75	1.14	0.4581
3.64	-0.87	0.5456	-1.09	1.48	0.3781	3.18	-0.83	0.5584
-2.26	1.62	0.3514	-0.66	4.41	0.1419	-2.91	0.79	0.5739
-7.01	3.56	0.1743	-0.89	6.41	0.0985 ^M	-1.12	18.04	0.0352*
-12.00	.	.	-1.91	1.61	0.3537	5.64	-0.71	0.6088
-2.94	2.24	0.2674	-0.92	2.61	0.2329	-2.12	3.74	0.1664
-3.39	1.31	0.4162	-0.79	5.9	0.1068	-0.27	9.63	0.0658 ^M
-40.00	.	.	-0.79	1.02	0.4923	-1.36	59.47	0.0107*
-1.18	2.12	0.2801	-2.40	.	.	-0.55	1.53	0.3686
3.00	.	.	-0.54	6.31	0.1000	-0.00	.	.
-0.00	.	.	-8.00	.	.	-0.07	0.03	0.9807
-0.00	.	.	-0.77	1.45	0.3842	-8.00	.	.
-1.63	49.65	0.0128*	-1.88	1.47	0.3798	-3.95	1.58	0.3600
1.05	-0.12	0.9268	-2.21	4.84	0.1298	-0.81	1.87	0.3126
-0.65	.	.	-2.69	1.67	0.3440	-0.57	3.08	0.1999
-0.13	0.07	0.9534	-0.54	5.94	0.1062	-0.97	3.05	0.2017
-2.64	20.4	0.0312*	-1.73	5.67	0.1112	-3.47	3.99	0.1564
-0.00	.	.	-1.15	1.97	0.2997	-0.69	0.59	0.6613
-2.78	2.89	0.2123	2.68	-0.3	0.8160	-6.67	2.89	0.2123
-2.86	.	.	-0.50	0.83	0.5573	-1.31	1.22	0.4380
120.00	.	.	-2.11	2.19	0.2727	-0.57	0.72	0.6047
-3.47	1.85	0.3154	-1.09	1.71	0.3376	-3.72	4.1	0.1523
-3.56	0.30	0.8143	-5.69	4.13	0.1512	-1.11	36.52	0.0174*
13.33	.	.	-0.55	3.45	0.1796	-1.00	14.64	0.0434*
			-0.69	3.72	0.1674	-1.78	.	.
			-1.79	4.04	0.1544	-1.24	3.61	0.1720
			-0.67	0.99	0.5030	-1.10	0.96	0.5122
			-0.47	3.09	0.1993			
			-0.40	0.33	0.7987			
			-0.78	27.67	0.0230*			
			-0.72	4.26	0.1468			
			-0.80	0.78	0.5769			
			-0.76	60.62	0.0105*			

Notes: p-values use a modified Michelin scale, adding marginal significance to the uncertainty region ($0.05 \leq p \leq 0.1$): * significant, ** highly significant, ^M marginally significant

The overall trend was very highly significantly decreasing ($p < 0.001$), with slight variations among subjects assigned to the classical intervention ($p = 0.025$), intensive intervention assisted by nutritionists ($p < 0.001$), and intensive intervention assisted by psychologists ($p < 0.001$). Comparisons among groups suggested that intensive intervention methods are more efficient with respect to weight loss than the classical ones ($p = 0.006$ when joining interventions assisted by nutritionists and psychologists, 0.027 otherwise), but no significant differences were detected between interventions assisted by nutritionists and psychologists ($p = 0.388$).

The results were checked for consistency with findings using a traditional approach, employing the Analysis of Variance (ANOVA) to test whether there are significant differences between the three groups. In this case, we defined the weight loss as the difference between the initial ($t = 0$) and final ($t = 12$) weight. The global F test was $F = 10.88$ with $p < 0.001$, suggesting the existence of significant differences between groups. No differences were found between interventions assisted by nutritionists and psychologists, but both of them differed significantly from the classical approach, being more efficient.

3.2. Example #2: Clinical meaning of echocardiographic parameters

Data were generated within the Acute Myocardial Reperfusion Syndrome study "BNP – prognostic value of BNP correlated with echocardiographic indices of systolic and diastolic function in patients with ST elevation acute myocardial infarction with indication of reperfusion" (National University Research Council grant #22 of 2006) looking at risk factors, predictors and criteria assessing the success of interventions.

In this case, the variables of interest were the ejection fraction, defined as the difference between end-diastolic and end-systolic volumes divided by end-diastolic volume [4], and the kinetic score, computed according to the guidelines of the American Society of Echocardiography as the ratio between the sum of scores assigned to each segment of the left ventricle (1 = normal; 2 = hypokinetic; 3 = akinetic; 4 = dyskinetic) and their number, i.e. 16 [5]. Both values were recorded in the beginning of the study (0 days), during the study, after 1, 7 and respectively 30 days, and in the end of the study (365 days). Even though 88 subjects were included in the study, values of the ejection fraction and kinetic score were not always recorded, and the actual sample size was diminished.

The sign test did not detect any trend with respect to the ejection fraction ($p = 0.198$), but detected a significantly decreasing trend of the kinetic score ($p = 0.003$).

However, in this particular study further research questions that could be answered using the proposed methodology. Among them, of particular interest was the predictive value for screening purposes of recording a value of either or both the ejection fraction and/or the kinetic score at one of the particular moments used (0, 1, 7, 30 or 365 days). In order to answer this question, we assigned ranks from 1 to 5 to the pair (ejection fraction, time) and (kinetic score, time) with a maximum impact on the regression line, corresponding to its position on the time scale: 1, for $t = 0$; 2, for $t = 1$; 3, for $t = 7$; 4, for $t = 30$; and 5, for $t = 365$.

The magnitude of the impact was assessed based on the jackknife residual [6]:

$$r_{(-i)} = \frac{y_i - (\alpha + \beta \cdot x_i)}{\sqrt{\sigma_{(-i)}^2 \left(1 - \frac{1}{n} + \frac{(x_i - \mu_x)^2}{\sum_{i=1}^n x_i^2 - n \mu_x^2} \right)}} \approx t_{n-3}$$

In situations where the jackknife residuals could not be computed, ranks were determined based on the value of the actual residuals [6]:

$$e_i = Y_i - (\alpha + \beta \cdot X_i)$$

The maximum impact corresponded to the maximum absolute value of either the jackknife residual or actual residual, within each set of five values computed for each patient.

In order to test whether some moment has a higher predictive value, we tested the statistical significance of the difference between the proportion of its corresponding rank and 0.2 (proportion under the null hypothesis), using a modified version of the test proposed by Pagano and Gavreau [8]:

$$z = \frac{|p-0.2|}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$$

Results suggest that for both variables the first moment (t=0) appears to be the most important, as its corresponding p-values are the lowest (Table 2). Nevertheless, for the ejection fraction none of the moment was significant at 0.05. Due to the reduced sample sizes, we accounted for additional marginal significance if p values were in the uncertainty area (0.05 ≤ p ≤ 0.1). In the same order of importance, the second moment is t=365 for the ejection fraction and t=30 for the kinetic score.

The usage of two variables in the second study allowed for checking the validity of the proposed approach. Since the ejection fraction and the kinetic score were very highly significantly correlated (R²=-0.74 overall, -0.62 at t=0, -0.71 at t=1, -0.75 at t=7, -0.74 at t=30 and -0.84 at t=365, with p<0.001 in all cases), we assessed the correlation at the subject level looking at the correlation of the ranks described above. Spearman's coefficient of correlation was 0.189 with p=0.086, falling into the uncertainty region. This could be due to reducing the overall sample as the computation of ranks was not always possible since five values were not always recorded for each variable for individual subjects.

Table 2. Tests of the predictive value of the moment when the ejection fraction and kinetic score are recorded: Bucharest, 2008

Ejection fraction (n=81)				Kinetic score (n=86)			
Rank	Frequency	z	p	Rank	Frequency	z	p
1	0.28	1.63	0.0516 ^M	1	0.32	2.33	0.0099 ^{**}
2	0.20	0.05	0.4801	2	0.15	1.31	0.0951 ^M
3	0.17	0.63	0.2643	3	0.11	2.55	0.0054 [*]
4	0.21	0.21	0.4168	4	0.12	2.09	0.0183 [*]
5	0.14	1.62	0.0526 ^M	5	0.30	1.90	0.0287 [*]

Notes: p-values use a modified Michelin scale, adding marginal significance to the uncertainty region (0.05 ≤ p ≤ 0.1): * significant, ** highly significant, ^M marginally significant

4. Conclusion

Both examples indicate that the proposed approach was able to produce significant results despite of the reduced sample sizes. Comparisons with classical approaches, answering partially the same research question, suggest that the proposed methodology is valid and could be used to answer the particular question of assessing subject-level trends is

clinical trials. The only disadvantage is that it employs a large number of analyses, problem resolved by employing appropriate software.

References

1. Bernard, Y. **L'échocardiogramme de stress**, Revue Médicale Suisse, no. 2258, June 2, 1999, online source: <http://revue.medhyg.ch/print.php3?sid=19884>, accessed on March 7, 2009
2. Gordis, L. **Epidemiology**, Philadelphia, PA: W. B. Saunders Company, 1996, pp. 90
3. Habash-Bseiso, D. E., Rokey, R., Berger, C. J., Weier, A. W. and Chyou, P.-H. **Accuracy of Noninvasive Ejection Fraction Measurement in a Large Community-Based Clinic**, Clinical Medicine and Research 3, no. 2, 2005, pp. 7-82
4. Hussey, J. R. **BIOS 757: Intermediate Biometrics. Course notes/Spring 1998**, Columbia, SC: School of Public Health, University of South Carolina, 1998, pp. 9,15, 29, 30
5. Lilienfeld, A. M. and Lilienfeld. D. E. **Foundations of Epidemiology**, 2nd ed., New York, NY: Oxford University Press, 1980, pp. 257-258
6. Mausner, J. S. and Kramer, S. **Mausner & Bahn Epidemiology – An Introductory Text**, Philadelphia, PA: W. B. Saunders Company, 1985, pp.195
7. Pagano, M. and Gauvreau, K. **Principles of Biostatistics**, 1st ed., Belmont, CA: Duxbury Press, 1993, pp. 296
8. Piegorsch, Walter W. and John A. Bailer **Statistics for Environmental Biology and Toxicology**, 1st ed., London, UK: Chapman and Hall, 1997, pp. 12

¹ Codification of references:

[1]	Gordis, L. Epidemiology , Philadelphia, PA: W. B. Saunders Company, 1996, pp. 90
[2]	Lilienfeld, A. M. and Lilienfeld. D. E. Foundations of Epidemiology , 2nd ed., New York, NY: Oxford University Press, 1980, pp. 257-258
[3]	Mausner, J. S. and Kramer, S. Mausner & Bahn Epidemiology – An Introductory Text , Philadelphia, PA: W. B. Saunders Company, 1985, pp.195
[4]	Habash-Bseiso, D. E., Rokey, R., Berger, C. J., Weier, A. W. and Chyou, P.-H. Accuracy of Noninvasive Ejection Fraction Measurement in a Large Community-Based Clinic , Clinical Medicine and Research 3, no. 2, 2005, pp. 7-82
[5]	Bernard, Y. L'échocardiogramme de stress , Revue Médicale Suisse, no. 2258, June 2, 1999, online source: http://revue.medhyg.ch/print.php3?sid=19884 , accessed on March 7, 2009
[6]	Hussey, J. R. BIOS 757: Intermediate Biometrics. Course notes/Spring 1998 , Columbia, SC: School of Public Health, University of South Carolina, 1998, pp. 9,15, 29, 30
[7]	Piegorsch, Walter W. and John A. Bailer Statistics for Environmental Biology and Toxicology , 1st ed., London, UK: Chapman and Hall, 1997, pp. 12
[8]	Pagano, M. and Gauvreau, K. Principles of Biostatistics , 1st ed., Belmont, CA: Duxbury Press, 1993, pp. 296

STATISTICAL MODELING OF THE INCIDENCE OF BREAST CANCER IN NWFP, PAKISTAN

Salah UDDIN

PhD, University Professor, Chairman, Department of Statistics,
University of Peshawar, Peshawar, NWFP, Pakistan

E-mail: salahuddin_90@yahoo.com

Arif ULLAH

Lecturer in Statistics, Higher Education Department, Peshawar, NWFP, Pakistan

E-mail:

NAJMA

Lecturer in Statistics, Frontier Women University, Peshawar, NWFP, Pakistan

E-mail:

Muhammad IQBAL

Lecturer, Department of Statistics,
University of Peshawar, Peshawar, NWFP, Pakistan

E-mail:

Abstract: Breast cancer is the most common form of cancer that affects women. It is a life threatening disease and the most common malignancy in women through out the world. In this study an effort has been made to determine the most likely risk factors of breast cancer and to select a parsimonious model of the incidence of breast cancer in women patients of the age 50 years and above in the population of North West Frontier Province (NWFP), Pakistan. The data were collected from a total of 331 women patients, arriving at Institute of Radiotherapy and Nuclear Medicine Peshawar, NWFP, Pakistan.

Logistic regression model was estimated, for breast cancer patients, through backward elimination procedure. Brown tests were applied to provide an initial model for backward elimination procedure. The logistic regression model, selected through backward elimination procedure contains the factors Menopausal status (M), Reproductive status (R), and the joint effect of Diet and family History (D*H). We conclude that menopausal status; reproductive status and the joint effect of diet and family history were the important risk factors for the breast cancer.

Separate models were then fitted for married and unmarried breast cancer patients. The best-selected model for married females is of factors Feeding (F), R, M, (D*H), whereas the best selected model for unmarried females has only one main factor Menopausal status. We conclude that breast feeding, reproductive status, menopausal status and the joint effect of diet and family history were the important risk factors of breast cancer in married women and the menopausal status was the important risk factor of breast cancer in unmarried women.

Key words: Logistic regression; backward elimination procedure;
Brown method; Wald statistic

1. Introduction

Cancer of breast is a disease that instills feelings of dread and fear in many women. Not only is it a life threatening disease, but it affects a part of the body that is central to women's sense of womanliness and femininity. It is a complex disease with the causes not yet fully understood. It is most likely caused by a number of factors interacting with each other, rather than by any one factor. The main identified risk factor of breast cancer is age, with woman aged 80 years or over being most at risk (Jelfs, 1999).

According to Australian Institute of Health and Welfare report (AIHW, 1998), 43% of breast cancer cases were in women between the ages of 45 and 64, and 22% were in women between the ages of 65 and 74 during 1982-1994. Approximately 18% of cases occurred in women younger than 45 years and women older than 74 years.

Before the age of 75, one in eleven women in Australia is expected to diagnose with breast cancer. In 1996, 9556 new cases of breast cancer were diagnosed in Australia and there were 2,623 deaths attributed to breast cancer. Similarly, of the 30201 deaths from breast cancer in Australian women from 1982-1994, 38% occurred in women aged 45 to 64, 28% in women aged 75 or over, 24% in women aged 65 to 74, and 10% in women younger than 45 years of age (Kricger and Jelfs, 1996).

For the period 1996-2000, women aged 20-24 have the lowest incidence rate, 1.4 cases per 100,000 population; women aged 75-79 have the highest incidence rate, 499 cases per 100,000 (Ries et al., 2003). Breast cancer incidence rates among African American women range from 89.8 in Rhode Island to 147.6 in Alaska (Hotes et al., 2003).

Until now no such statistical study has been made in the province of NWFP on the various risk factors of breast cancer. In this study an effort has been made to model the relationship between breast cancer in the population of NWFP and its probable risk factors.

2. Data Set

The data for this study were collected from Institute of Radiotherapy and Nuclear Medicine (IRNUM), Peshawar. The study is based on a sample of size 331 women, including 123 cases (breast cancer patients) and 208 control (not breast cancer patients) groups. Out of 331 women breast cancer patients, 31 (9.37%) patients are unmarried and 300 (90.63%) patients are married.

The suggested risk factors for fitting the model are family history (H), reproductive status (R), breast-feeding (F), oral contraceptives (C), menopausal status (M) and diet (D). The response variable for the study is the diagnosis of patient with breast cancer or not.

3. Method and Materials

Generalized linear models introduced by Nelder and Wedderburn (1972) are a class of statistical models, which is the natural generalization of classical linear model. It includes response variables that follow any probability distribution in the exponential family of distributions. An excellent treatment of generalized linear models is presented in Agresti (1996). In this study the response variable is binary; therefore, the logistic regression model is an appropriate model, which is a part of generalized linear models.

The response variable in logistic regression is usually dichotomous, that is, the response variable can take the value 1 with a probability of success θ , or the value 0 with probability of failure $1-\theta$. This type of variable is called a Bernoulli (or binary) variable.

The relationship between the predictor and response variables is not a linear function in logistic regression; instead, the logistic regression function is used, which is given as

$$\theta(x) = \frac{e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}{1 + e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}} \quad (1)$$

We now find the link function for which the logistic regression model is a generalized linear model (GLM). For this model the odds of making response 1 are

$$\frac{\theta(x)}{1 - \theta(x)} = e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)} \quad (2)$$

$$\log \left[\frac{\theta(x)}{1 - \theta(x)} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (3)$$

Thus the appropriate link is the log odds transformation, the logit. The logistic regression model is given by

$$\text{logit} [\theta(x)] = \log \left[\frac{\theta(x)}{1 - \theta(x)} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (4)$$

The parameters in this model, $\alpha, \beta_1, \beta_2, \dots, \beta_k$ can no longer be estimated by least squares, but are found using the maximum likelihood method (Collett, 1991; Cox & Snell, 1989).

Logistic regression calculates the probability of success over the probability of failure; therefore, the results of the analysis are in the form of an odds ratio. Logistic regression also provides knowledge of the relationships and strengths among the variables.

The Wald statistic is commonly used to test the significance of individual logistic regression coefficients for each independent variable. The Wald statistic for the β_j coefficient is:

$$\text{Wald} = \left[\frac{\hat{\beta}_j}{S.E.(\hat{\beta}_j)} \right]^2,$$

It is distributed as chi-square with 1 degree of freedom. The Wald statistic is simply the square of the (asymptotic) t -statistic. The Wald statistic can be used to calculate a confidence interval for β_j . We can assert with $100(1-\alpha)\%$ confidence that the true parameter lies in the interval with boundaries $\hat{\beta} \pm Z_{\alpha/2}(ASE)$, where ASE is the asymptotic standard error of logistic $\hat{\beta}$. Parameter estimates are obtained using the principle of maximum likelihood; therefore hypothesis tests are based on comparisons of likelihoods or the deviances of nested models. The likelihood ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The likelihood-ratio test statistic equals:

$$-2 \log \left(\frac{L_0}{L_1} \right) = -2 [\log(L_0) - \log(L_1)] = -2(L_0 - L_1) \quad (5)$$

This log transformation of the likelihood functions yields a chi-squared statistic. This is the recommended test statistic to use when building a model through backward elimination procedure. Once $\hat{\beta}$ has been obtained, the estimated value of the linear systematic component (also known as linear predictor) of the model is

$$\hat{\eta}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki} \quad (6)$$

From equation (6), the fitted probabilities $\hat{\theta}_i$ can be found using

$$\hat{\theta}_i = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}}$$

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. Several different options are available during model creation. Variables can be entered into the model in the order specified by the researcher or logistic regression can test the fit of the model after each coefficient is added or deleted (backward elimination procedure), called stepwise procedure. Backward elimination procedure appears to be the preferred method of exploratory analyses, where the analysis begins with a full or saturated model and variables are eliminated from the model in an iterative process. The fit of the model is tested after the elimination of each variable to ensure that the model still adequately fits the data. When no more variables can be eliminated from the model, the analysis has been completed.

4. Analyses and Interpretation

We begin with the initial model having factors F, M, R, (F*D), (M*C), (D*C), (D*H), and (F*D*R), provided by Brown test. Using backward elimination method through SPSS package, the final model was selected at step 6 which contains two main factors (M and R) and one interaction factor (D*H). Thus the significant factors are M, R, (D*H). It means that menopausal status (M), reproductive status (R) and the joint effect of diet and family history (D*H) were the important risk factors for the breast cancer.

Table 1. Variables in Model 1

Variables	$\hat{\beta}$	S.E($\hat{\beta}$)	Wald	d.f.	P-value	Exp($\hat{\beta}$)	95% C.I. for Exp($\hat{\beta}$)	
							Lower	Upper
(D*H)	2.235	0.820	7.432	1	0.006	9.350	1.874	46.644
M	2.449	0.472	26.907	1	0.000	11.576	4.589	29.203
R	1.298	0.328	15.608	1	0.000	3.660	1.923	6.967
Constant	-1.083	0.147	54.361	1	0.000	0.339	-	-

The fitted model is:

$$\text{Logit}(\hat{p}) = -1.083 + 2.449M + 1.298R + 2.235(D*H) \quad (7)$$

4.1 Analysis According to Marital Status

Some factors like reproductive status (R), breast-feeding (F) and oral contraceptives (C) are irrelevant to unmarried women. Therefore, Separate models are fitted for married and unmarried women patients.

(a) Model for Married Women

The suggested risk factors (explanatory variables) in this case are family history (H), reproductive status (R), breast-feeding (F), oral contraceptives (C), menopausal status (M) and diet (D). We repeat the same process, starting from Brown method and then backward elimination procedure. The final model selected at step 5, contains the factors F, M and (D*H). It means that one new factor breast-feeding (F) is turned out significant in this case and other significant factors M, (D*H) are the same.

Table 2. Variables in Model 2

Variable	$\hat{\beta}$	S.E($\hat{\beta}$)	Wald	d.f	P-value	Exp($\hat{\beta}$)	95% C.I. for Exp ($\hat{\beta}$)	
							Lower	Upper
(D*H)	1.746	0.843	4.001	1	0.045	5.730	1.036	31.701
M	2.382	0.525	20.574	1	0.000	10.823	3.867	30.291
F	1.342	0.368	13.287	1	0.000	3.827	1.860	7.875
Constant	-1.129	0.170	44.300	1	0.000	0.323	-	-

The fitted model is:

$$\text{Logit}(\hat{p}) = -1.129 + 1.342F + 2.382M + 1.746 (D*H) \tag{8}$$

(b) Model for Unmarried Women

The suggested risk factors in this case are diet (D), menopausal status (M) and family history (H). We begin with the initial model having factors D, M & H, provided by Brown method for backward elimination method. The procedure selects the final model at step 1 with only one main factor M. Hence the menopausal status (M) is the important risk factor of breast cancer in unmarried case.

Table 3. Variables in Model 3

Variable	$\hat{\beta}$	S.E($\hat{\beta}$)	Wald	d.f	P-value	Exp($\hat{\beta}$)	95% C.I. for Exp ($\hat{\beta}$)	
							Lower	Upper
M	2.748	1.090	6.354	1	0.012	15.610	1.843	132.222
Constant	-0.671	0.256	6.856	1	0.009	0.511	-	-

The fitted model is

$$\text{Logit}(\hat{p}) = - 0.671 + 2.748 (M) \tag{9}$$

In this model the variable 'M' is selected as an important factor. This model is the reduce form of model 1 and model 2.

5. Conclusion

The purpose of this study was to estimate a model to determine the most likely risk factors of breast cancer, women patients of age 50 years or above, in NWFP. The phenomena of breast cancer was studied in relation to different risk factors namely, oral contraceptives (C), diet (D), menopausal status (M), family history (H), breast feeding (F) and reproductive status (R) of 331 patients arriving at IRNUM Peshawar. Out of these 331 patients, the number of breast cancer patients (case group) were 123 and 208 had no breast cancer (control group); 31 (9.37%) were unmarried and 300 (90.63%) were married. Older women were at increasing risk over time.

For model fitting procedure we used a binary response variable B (Breast cancer), taking the value 1 for breast cancer patients and 0 otherwise. Brown method was used for selecting initial model. Backward elimination procedure was used to determine a parsimonious model. The logistic regression analysis was then applied to the data.

The variables chosen initially as predictors were R, H, C, D, M and F. The logistic regression model, selected through backward elimination procedure contains the factors M, R and the interaction term (D*H). It means that menopausal status; reproductive status and the joint factor (D*H) were the important risk factors for the breast cancer. Separate logistic regression model was then fitted for married and unmarried women patients. For married females, we obtained the model with predictors F, M, (D*H). Thus breast feeding, menopausal status and (D*H) were the important risk factors. For unmarried women on the other hand, we obtained the final model containing only one main factor M. Hence the menopausal status was the only important risk factor.

Finally on the basis of analysis based on sample of 331 patients, we concluded that menopausal status, reproductive status and joint effect of diet and family history are the important risk factors. While in addition to these factors, breast-feeding is also the most important factor in the case of married patients. Therefore, one of our main findings is that breast-feeding is a preventive measure against breast cancer. Furthermore, the risk of breast cancer is found to be increasing with age. The highest incidents among women were in the age group (40-59). Therefore, it is suggested that breast cancer screening may be advised before this age group.

References

1. Agresti, A. **An Introduction to Categorical Data Analysis**, John Wiley and Sons Inc., New York, 1996
2. AIHW **Breast Cancer Survival in Australian Women 1982-1994**, Canberra, Australian Institute of Health and Welfare, 1998, online source: www.aihw.gov.au/publications/health/bcsaw82-94/bcsaw82-94.pdf
3. Collett, D. **Modeling Binary Data**, Chapman & Hall, London, 1991
4. Cox, D. R. and Snell, E. J. **Analysis of Binary Data**, 2nd edition, Chapman and Hall, London, 1989
5. Hotes, J. L., McLaughlin, C. C., Lake, A., Frith, R., Roney, D., Cormier, M., Eulton, J. P., Holowaty, E., Howe, H. L., Kosary, C. and Chen, R. W. (eds.), **Cancer in North American 1996-2000, Volume No. 1: Incidence, Volume No. 2: Mortality**, Springfield IL, and North American Association of Central Cancer Registries, 2003
6. Jelfs, P. **Breast cancer in Australia: an overview, The Breast Cancer Bulletin 1999**, 1999, pp. 1-2

7. Kricke, A. and Jelfs, P. **Breast Cancer in Australian Women 1921-1994**, Canberra, Australian Institute of Health and Welfare, 1996, pp. 12
8. Nelder, J. A. and Wedderburn, R. W. M. **Generalized Linear Models**, Journal of Royal Statistical Society, Series A, 135, 1972, pp. 370-384
9. Ries, L. A. G., Eisner, M.P. and Kosary, C. L. (eds.) **SEER Cancer Statistics Review, 1975-2000**, Bethesda, MD: National Cancer Institute, 2003

AN EMPIRICAL INVESTIGATION OF META-ANALYSIS USING RANDOMIZED CONTROLLED CLINICAL TRIALS IN A PARTICULAR CENTRE

C. PONNURAJA¹

Department of Statistics,
Tuberculosis Research centre (ICMR), Chennai, India

E-mail: cponnuraja@gmail.com

Abstract: *Meta-analysis is the combination of results from various independent studies. In a meta-analysis, combining survival data from different clinical trials, an important issue is the possible heterogeneity between trials. Such inter-trial variation can not only be explained by heterogeneity of treatment effects across trials but also by heterogeneity of their baseline risk. In addition, one might examine the relationship between magnitude of the treatment effect and the underlying risk of the patients in the different trials. However, the need for medical research and clinical practice to be based on the totality of relevant and sound evidence has been increasingly recognized. In this paper, we review the advances of meta-analysis using clinical trials TB data. This paper examines sixteen reporting results of randomized clinical trials conducted in a particular centre at consecutive periods. Every study pools that the results from the relevant trials in order to evaluate the efficacy of a certain treatment between cases and control. There is a need for empirical effort comparing random effects model with the fixed effects model in the calculation of a pooled relative risk in the meta-analysis in systematic reviews of randomized controlled clinical trials. We review heterogeneity and random effects analyses and assessing bias within and across studies. We compare the two approaches with regards to statistical significance, summary relative risk, and confidence intervals.*

Key words: *fixed effects model; random effects model; heterogeneity of treatment effects*

1. Introduction

Meta-analysis provides an objective way of combining information from independent studies looking at the same clinical questions and has been applied most often to treatment effects in randomized clinical trials. We understand meta-analysis as being the use of statistical techniques to combine the results of studies addressing the same question into a summary measure. Standard meta-analysis methods for providing an overall estimate of the treatment effects rely on certain assumption (Whitehead and Whitehead, 1991). Meta-analysis is the term given to retrospective investigations in which data from all known studies of a particular clinical issue are assembled and evaluated collectively and quantitatively. It differs in important ways from traditional narrative reviews, in that there is a commitment to scientific principles in assembling and analyzing the data, via protocol-driven library searches and data abstraction, in addition to the formalism of statistical analysis. There is a

need for more empirical work on methodology, properties and limitations of underlying statistical methodology (Engels, et al, 2000). Heterogeneity, by which we mean variation among the results of individual trials beyond that expected from chance alone, is an important issue in meta-analysis. Heterogeneity may indicate that trials evaluated different interventions or different populations. It is clear that when there are substantial differences among trial results, and in the face of heterogeneity, a single estimate may be misleading and should be avoided and exploration of heterogeneity is also a critical important component of meta-analysis of randomized trials (Thompson and Pocock, 1991; Thompson 1994; Lau et al. 1995). Most of the arguments presented against random effects model could be considered as explanations of the limitations of using covariates to explain the heterogeneity in trial results. There is limited empirical experience comparing results from random effects and fixed effects models, particularly when the results are heterogeneous (Thompson and Pocock, 1991). The random effects model incorporates the heterogeneity of treatment effects across studies in the analysis of the overall treatment efficacy (DerSimonian and Liard, 1986). We present an empirical investigation from meta-analysis of randomized clinical trials included in systematic reviews as well as reports conducted in the area of tuberculosis infected patients; we compare the two approaches with regards to statistical significance, summary relative risk, and confidence intervals. The results of any individual trial must be absorbed and debated by the scientific community before wholesale recommendations regarding treatment practice are observed. Randomized trials and meta-analyses have distinct but complementary goals. Meta-analysis can be used productively in planning new clinical trials, and in supplying updated information to study monitors in the course of a trial. This process of debate necessarily involves the weighing of evidence from different sources, and meta-analysis can and does play an important role in this process (Begg, 1996).

2. Definition of models

The two models have been used here, they are fixed effects model (FEM) and random effects model (REM). Fixed effects model assumes that there is a common effect and a random component, which means sampling error, is responsible for difference among trial results, that is, it assumes heterogeneity of intervention effects. This approach provides inferences only about the set of trials under review, giving weight to each trial based on the 'within study' sampling variance. The individual study sample size and the number of events are the leading factors in the weight assigned to each trial in the pooled estimate of the relative risk. The FEM formulations are inverse variance method, Mantel-Haenszel method and Peto's method. However, the Peto's modified estimate can give biased answers in a few circumstances, such as when there is severe imbalance in treatment allocation within individual studies or in the presence of very large treatment effects. The REM provides inference based on the assumption that the observed trials are a sample from a hypothetical population of trials. Also to account for the variation among trials results a random term is added to compute the weights in the REM, representing 'among' trials variation, as often estimated from a function of the chi-squared test for heterogeneity. This term adds a common variance component to the weight of each trial in the meta-analysis, which tends equalize the weights assigned to small and large trials (Villar, et al., 2001). The disproportionate overall influence of small trials is more evident when there is heterogeneity

of trial results because the 'among' trials variance becomes larger and dominates the within-trial random effects.

When heterogeneity is present, it may be inappropriate to combine the separate trial estimates into a single number, particularly using fixed effects methods that assume a common treatment effect. Random effects methods, which provide an attractive approach to summarizing heterogeneous results, model heterogeneity as variation of individual trial treatment effects around a population average effect. The key distinction between these two types of models concerns the belief regarding behavior of trial effects as trial sample sizes get very large. If one believes that the individual trial effects would converge to a common value for all trials, a fixed effects model is appropriate, whereas if one believes that individual trials would still demonstrate separate effects, then a random effects model is preferable (Thompson and Pocock, 1991). The random effects model anticipates better than the fixed effects model by Fleiss (1993) and also the National Research Council (1992) make known the benefits of using random effects model.

3. A meta-analysis of sixteen randomized clinical trials

For the present analysis we examine sixteen clinical trials at same centre each reporting results from several independent trials over a period between 1956 and 1995. All the sixteen trials have been categorized into two groups based on their duration segment. Each review pools the results from the relevant trials in order to evaluate the efficacy of a certain treatment for a specified condition. These reviews lack of consistent assessment of homogeneity of treatment effect before pooling. We discuss both fixed effects and random effects approach to combining evidence from a series of experiments comparing two treatments. This approach incorporates the heterogeneity of effects in the analysis of the overall treatment efficacy. The model can be extended to include relevant covariates which would reduce the heterogeneity and allow for more specific therapeutic recommendations. Most often to explore heterogeneity is stratification. Studies are categorized according to the characteristics of the study or the characteristics of the subjects in the study and a summary estimate of effect is estimated in each of the categories (Petitti, 2001).

4. Statistical methods

Results of the outcome were abstracted and are expressed as summary relative risk and 95 per cent confidence interval (CI) for both random and fixed effects models. The summary relative risk for the FEM was calculated using the Mantel-Haenszel method while the DerSimonian and Laird method was used for the REM.

Mantel-Haenszel Method

This is for calculating a summary estimate of effect across strata. Since studies are identified for a meta-analysis as strata, the Mantel-Haenszel method is an appropriate for analyzing data for a meta-analysis based on fixed effect. It is used when the measure of effect is a ratio measure. Kleinbaum, Kupper, and Morgenstern (1982) give formulas that would allow in Mantel-Haenszel to be applied. Notations for applications of Mantel-Haenszel

	Treated	Control	Total
Recurrent	a_i	b_i	g_i
Non Recurrent	c_i	d_i	h_i
Total	e_i	f_i	n_i

Summary odds ratio

$$OR_{mh} = \frac{Sum(W_i \times OR_i)}{Sum W_i}$$

$$OR_i = \frac{(a_i \times d_i)}{(b_i \times c_i)}$$

$$W_i = 1/\text{variance}_i$$

$$\text{Variance}_i = \frac{n_i}{(b_i \times c_i)}$$

$$95\% \text{ confidence interval} = e^{\ln OR_{mh}} \pm 1.96\sqrt{\text{variance} OR_{mh}}$$

where variance OR_{mh} is calculated as Robins, Greenland, and Breslow(1986). The

$$\text{Variance}_{mh} = \left(\frac{Sum F}{2 \times (Sum R)^2} \right) + \left(\frac{Sum G}{2 \times Sum R \times Sum S} \right) + \left(\frac{Sum H}{2 \times (Sum S)^2} \right)$$

where $F = a_i \times d_i \times \left(\frac{a_i + d_i}{n_i^2} \right)$

$$G = \frac{[a_i + d_i \times (b_i + c_i)] + [b_i \times c_i \times (a_i + d_i)]}{n_i^2}$$

$$H = \frac{[b_i \times c_i \times (b_i + c_i)]}{n_i^2}$$

$$R = \frac{a_i \times d_i}{n_i}$$

$$S = \frac{b_i \times c_i}{n_i}$$

Formula for calculate a statistic for a test of homogeneity of effects;

$$Q = Sum[W_i \times (\ln OR_{mh} - \ln OR_i)^2], \text{ where, } Q \text{ is referred to the chi-square distribution with}$$

one degree of freedom.

DerSimonian & Laird Method

The DerSimonian and Laird (1986) method is based on the random-effects model. Formulas for applying the DerSimonian-Laird method summarizing studies in the case where effects are measured as odds ratios are given by Fleiss and Gross (1991)

$$\ln OR_{dl} = \frac{Sum(W_i^* \times \ln OR_i)}{Sum(W_i^*)};$$

where OR_{dl} is the DerSimonian-Laird summary estimate of the odds ratio, W_i^* is the DerSimonian-Laird weighting factor for the i^{th} study, and OR_i is the odds ratio from the i^{th} study

$$W_i^* = \frac{1}{D + \left(\frac{1}{W_i}\right)} \quad \text{where } W_i \text{ is given in MH and}$$

$$D = \frac{[Q - (S - 1)] \times \text{Sum}W_i}{[(\text{Sum}W_i)^2 - \text{Sum}(W_i^2)]}; \text{ and } D = 0 \text{ if } Q < S - 1;$$

where **S** is the number of studies and

$$Q = \text{Sum}W_i (\ln OR_i - \ln OR_{mh})^2 \text{ from this formula}$$

$$95\% \text{ CI} = e^{\ln OR_{mh}} \pm 1.96 \sqrt{\text{variance}_i^*}; \text{ where}$$

$$\text{Variance}_i^* = \frac{1}{\text{Sum}W_i^*}$$

The fixed effects let **Y** denote the generic measure of the effect of an experimental intervention; let **W** denotes the reciprocal of the variance of effect size. Under the assumption of the fixed set of studies, an estimator of the assumed common underlying effect size is

$$\bar{Y} = \frac{\sum_{i=1}^N W_i Y_i}{\sum_{i=1}^N W_i}$$

and the standard error of the estimator is

$$SE(\bar{Y}) = \left[\sum_{i=1}^N W_i \right]^{-1/2}$$

let ψ is the population effect size for an approximate 100(1- α)% confidence interval, then

$$\bar{Y} - z_{\alpha/2} \sqrt{\sum_{i=1}^N W_i} \leq \psi \leq \bar{Y} + z_{\alpha/2} \sqrt{\sum_{i=1}^N W_i}$$

Under the assumption of random effects, the studies are random samples from a largest population, the mean population size $\bar{\psi}$, about which the study-specific effect size vary. An approximation 100(1- α) % confidence interval for $\bar{\psi}$ is

$$\bar{Y}^* - z_{\alpha/2} \sqrt{\sum_{i=1}^N W_i^*} \leq \bar{\psi} \leq \bar{Y}^* + z_{\alpha/2} \sqrt{\sum_{i=1}^N W_i^*}$$

where $W_i^* = (D + W_i^{-1})^{-1}$

$$\bar{Y}^* = \frac{\sum_{i=1}^N W_i^* Y_i}{\sum_{i=1}^N W_i^*}$$

D denotes the study variation in effect size and this is calculated as

$$D = 0 \text{ if } Q \leq N - 1$$

$$D = [Q - (N - 1)]/U \text{ if } Q > N - 1$$

as Der Simonian and Laird, (1986)

$$U = (N - 1) \left[\bar{W} - \frac{S_w^2}{N\bar{W}} \right]$$

where \bar{W} and S_w^2 are the mean and variance of the W_s

The inconsistency of studies are being measured based on the classical measure of heterogeneity is Cochran's Q, which is calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies, with the weights being those used in the pooling method. Q is distributed as a chi-square statistic with k-1 (number of studies minus one) degrees of freedom. Q has low power as a comprehensive test of heterogeneity (Gavaghan et al. 2000) in particular when the number of trials is small in meta-analysis. If the number of studies are large where Q has more power as a test of heterogeneity (Higgins et al. 2003). Q is included in each meta-analysis function because it forms part of the DerSimonian-Laird random effects pooling method (DerSimonian and Laird 1986). An additional test, due to (Breslow and Day 1980), is provided with the odds ratio meta-analysis. We transformed the summary relative risks and the corresponding upper and lower limits of the 95 per cent CI for the two models to the natural logarithmic scale. I-squared statistic describes the percentage of variation across studies that are due to heterogeneity rather than chance (Higgins and Thompson, 2002; Higgins et al., 2003).

$$I^2 = 100\% \times \frac{(Q - df)}{Q}$$

We calculated the mean and standard deviation and range of the summary relative risk obtained using the two methods. To assess the differences between the summary relative risks and between the widths of the Confidence Intervals obtained using the two methods we calculated the mean of the paired differences. To investigate the average relative risk as a function of the difference we plotted the differences between the logs of the relative risks (log RR random-log RR fixed) against the mean of these two values. Graphs were plotted separately by heterogeneity status. The statistical evaluation of bias was conducted using the Begg and the Egger test. The complete analysis performed by STATA version 9.1, the meta command uses inverse-variance weighing to calculate fixed and random effects summary estimates, and, optionally to produce a forest plot. The advantage in using Meta command is that we require variables containing the effect estimate and its corresponding standard error for each study. When one arm of a study contains no events- or, equally, all events - we have what is termed a "zero cell" in the 2 x 2 table. Zero cells create problems in the computation of ratio measures of treatment effect, and the standard error of either difference or ratio measures. If no relapses any of the trial of any one group, the estimated odds ratio is zero and the standard error cannot be estimated. A common way to deal with this problem is to add 0.5 to each cell of the 2 x 2 for the trial (Cox and Snell, 1989). Because our inclusion criteria selected meta-analyses that had few trials with arms with zero events, this correction for zero cells had a minimal impact on conclusions. If there are no events in either the intervention or control arms of the trial, however, then any measure of effect summarized as a ratio is undefined, and unless the absolute risk difference scale is used instead, the trial has to be discarded from the meta-analysis.

5. Results

The following table gives data from 16 randomized controlled clinical trials of tuberculosis patients consists of both long term and short term treatments. The effects of treatment are being compared based on fixed and random effects method using meta-analysis. Table1 shows the trials consists both experimental as well as control groups for treating the patients. .

Table 1. Summary trials' data

Study name	Study year	Treated Group			Control Group		
		Total	Cured	Relapse	Total	Cured	Relapse
STNO1	1956	82	67	5	81	72	7
STNO3	1957	216	133	8	86	78	10
STNO5A	1961	72	68	5	66	56	7
STNO5B	1962	128	96	9	66	54	2
STNO7	1963	279	216	19	96	91	18
STNO8	1967	170	148	18	176	150	15
STNO9	1968	83	72	3	90	79	2
STNO10	1970	211	177	76	205	189	38
STNO11	1972	82	69	5	87	69	3
STNO11A	1973	86	74	1	87	76	2
STNO12	1974	261	261	24	269	269	24
STNO13	1977	228	219	42	466	257	64
STNO14	1980	111	111	3	117	117	7
STNO16	1986	305	294	15	512	495	52
STNO17	1990	594	562	25	273	259	16
STNO18	1995	184	182	15	176	174	9

The table 2 shows the magnitude of the change in the pooled estimate given by the random and fixed effects models to the trials between long-term treatment trials, short-term treatment trials and their combination in the calculation of the meta-analysis (exponential form) of tuberculosis care for infected individuals.

Table 2. The magnitude of the change in the pooled estimate

Trials	N	Pooled estimate in the meta-analysis		Test of Heterogeneity		No. of Trials in meta-analysis	Moment -based estimate of studies Variance
		REM	FEM	Q statistic	P value		
Long Term	2449	0.985	1.156	21.6 (9df)	P<0.05	10	0.325
Short Term	3496	0.778	0.774	6.6 (5df)	P>0.05	6	0.036
Combined	5955	0.193	0.251	45.3 (15df)	P<0.001	16	0.313

The tests of the heterogeneity are statistically significant in long-term trials and combined trials of long-term and short-term. Even though it is arguably sufficient, not possible to examine the null hypothesis that all studies are evaluating almost same effect

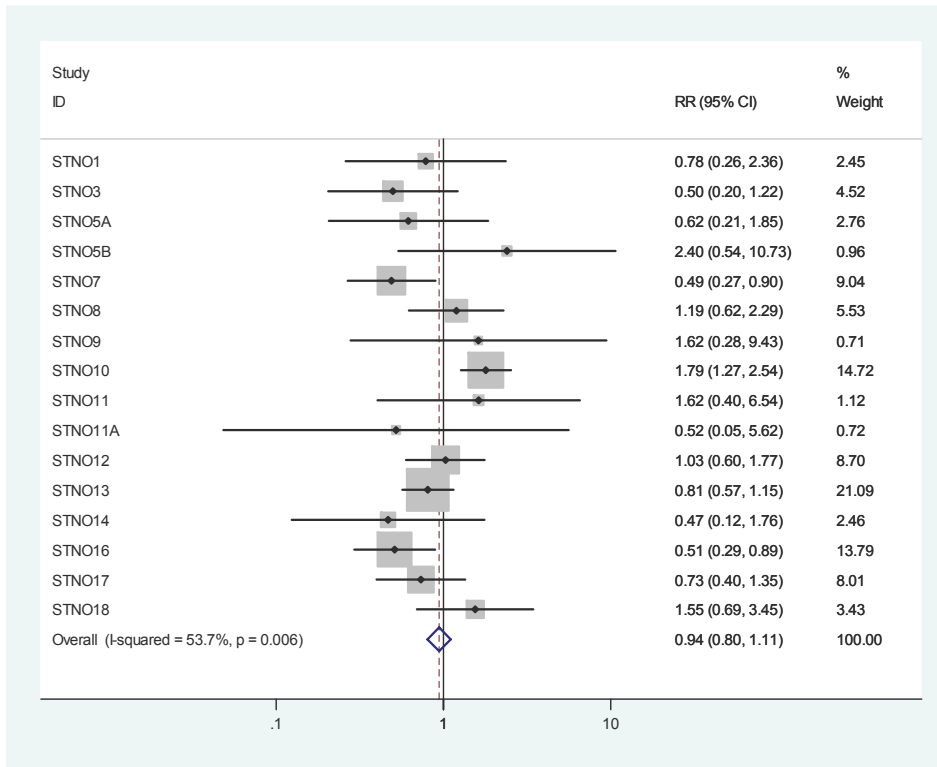


Figure 1. Forest Plot

In a forest plot the contribution of each study to the meta-analysis (its weight) is represented by the area of a box whose centre represents the size of the treatment effects estimated from that study. The summary treatment effect is shown by a middle of a diamond whose left and right extremes represent the corresponding confidence interval. Both the output and the graph show that there is a clear effect of treatments curing tuberculosis among patients. The meta-analysis dominated by the large study13, study10 and study16 trials which contribute around 50% of the weight in this analysis. Moreover the I-squared is constructed the inconsistency is 53.7 % (P=0.006).

Table 3. The summary of treatment effect

Study	Weights		Est	95% CI	
	Fixed	Random		Lower	Upper
STNO1	2.69	1.46	0.17	0.05	0.57
STNO3	4.08	1.79	0.19	0.07	0.50
STNO5A	2.66	1.45	0.20	0.06	0.66
STNO5B	1.56	1.05	0.13	0.03	0.63
STNO7	8.08	2.29	0.29	0.14	0.57
STNO8	7.37	2.23	0.22	0.11	0.46
STNO9	1.16	0.85	0.07	0.01	0.41
STNO10	19.84	2.75	0.63	0.41	0.98
STNO11	1.78	1.14	0.13	0.03	0.50
STNO11A	0.66	0.54	0.04	0.00	0.45
STNO12	11.00	7.86	1.03	0.57	1.86
STNO13	20.88	11.86	0.77	0.50	1.18
STNO14	2.03	1.89	0.45	0.11	1.79
STNO16	10.95	7.83	0.49	0.27	0.88
STNO17	9.25	6.92	0.72	0.38	1.37
STNO18	5.29	4.44	1.59	0.68	3.74

Note that remarkable differences between the fixed and random effects summary estimates in the long term and the combination of long term and short term trials, which arises because the studies are weighted much more equally in the random effects analysis. This shows the accountability of heterogeneity is comparable more in random effects than in the fixed effects method. Figure 2 based on random effects, shows the overall performances both fixed and random effects analyses. It is clear that the smaller studies such as study 12 and study 13 are given relatively more weight in the random effects than with the fixed effect model.

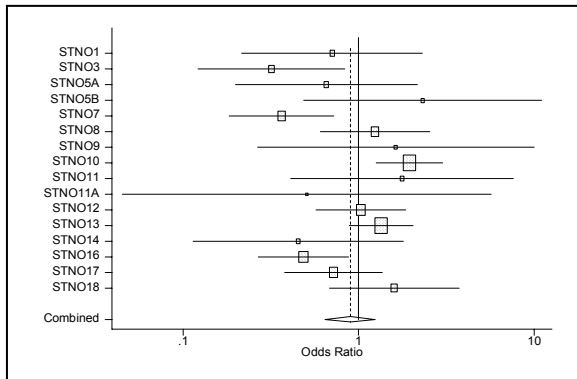


Figure2a. Forest Plot

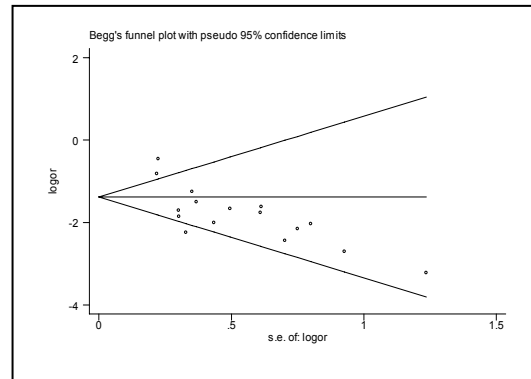


Figure2a. Funnel Plot

The method of assessing the effect of bias is using funnel plot as given below. In which the effect sizes from a study is plotted against the study's sample size. There is evidence of bias using the Eggar test based on weighted regression method ($p=0.004$) but not using the Begg such as rank correlation method. It is assuming that there is no heterogeneity but here there are three studies are significantly differing due to heterogeneity.

6. Discussions

The two approaches, the assumptions of a fixed and random set communicate the basis of estimation for each approach for a general measure of effect size. The fixed effect model is conditional on the stronger assumption that there is no true heterogeneity between studies also they are all estimating the same true effect and only differ because of sampling variation, where as the random effects method attempts to incorporate statistical heterogeneity into overall estimate of an average effect. The random effects model predicts better than the fixed effects model also to conclude that the modeling would be improved by an increase in use of random effects model than the fixed effects model. There is reviews focused meta-analysis using reviewed articles or published materials over a period or even in the several fields. But here we illustrated the meta-analysis applied for clinical trials in a particular centre and embossed the less heterogeneity among all the independent trials.

References

1. Begg. C. B. **The role of meta-analysis in monitoring clinical trials**, *Statistics in Medicine*, 15, 1996, pp. 1299-1306
2. Breslow, N. E. and Day, N. E. **Combination of results from a series of 2x2 tables; control of confounding**, In *Statistical Methods in Cancer Research, Volume 1: The Analysis of*

- Case-control Data. IARC Scientific Publications No. 32, International Agency for Health Research on Cancer: Lyon, 1980
3. Cox, D. R. and Snell, E. J. **Analysis of binary data**, Chapman and Hall, New York, 1989
 4. DerSimonian, R. and Laird, N. **Meta-Analysis in clinical trials**, Controlled Clinical Trials, 7, 1986, pp.177-188
 5. Egger, M., Zellwegar-Zahner, T., Schneider, M., Junker, C., Lengeler, C. and Antes, G. **Language bias in randomized controlled trials published in English and German**, Lancet, 350, 1997, pp. 326-329
 6. Engels, E., Schmid, C. H., Terrin, N., Olkin, I. and Lau, J. **Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses**, Statistics in Medicine, 19, 2000, pp. 1707-1728
 7. Everitt, B. S. and Pickles, A. **Statistical Aspects of the design and analysis of clinical trials**, Revised edition, Imperial College Press, London, 2004
 8. Fleiss, J. L. and Gross, A. J. **Meta-Analysis in Epidemiology, with special reference to studies of the association to environmental tobacco smoke and lung cancer: a critique**, Journal of Clinical Epidemiology, 44, 1991, pp.127-139
 9. Fleiss, J. L. **The statistical basis of meta-analysis**, Statistical Methods in Medical Research, 2, 1993, pp. 121-145
 10. Gavaghan, D. J. , Moore, A. R. and McQay, H. J. **An evaluation of homogeneity tests in meta-analysis in pain using simulations of patient data**, Pain, 85, 2000, pp. 415-424
 11. Higgins, J. P. T. and Thompson, S. G. **Quantifying heterogeneity in a meta-analysis**, Statistics in Medicine, 21, 2002, pp.1539-1558
 12. Kleinbaum, D. G., Kupper, L. L. and Morgenstern, H. **Epidemiologic Research: Principles and Quantitative Methods**, Belmont, Calif, Lifetime Learning, 1982
 13. Lau, J., Schmid, C. H. and Chalmers, T. C. **Cumulative meta-analysis of clinical trials build evidence for exemplary meta-analysis**, Journal of Clinical Epidemiology, 48, 1995, pp. 45-57
 14. Mantel, N. and Haenszel, W. **Statistical aspects of the analysis of data from retrospective studies in disease**, Journal of the National Cancer Institute, 22, 1959, pp. 719-748
 15. Petitti, D. B. **Approaches to heterogeneity in meta-analysis**, Statistics in Medicine, 20, 2001, pp. 3625-3633
 16. Robins, J., Greenland, S. and Breslow, N. E. **A general estimator for the variance of the Mantel- Haenszel odds ratio**, American Journal of Epidemiology, 124, 1986, pp. 719-723
 17. Thompson, S. G. and Pocock, S. **Can meta-analysis be trusted?**, Lancet, 338, 1991, pp. 1127-1130
 18. Thompson, S. G. **Why sources of heterogeneity in meta-analysis should be investigated**, British Medical Journal, 309, 1994, pp.1351-1355
 19. Villar, J., Mackey, M. E., Carroli, G. and Donnar, A. **Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models**, Statistics in Medicine, 20, 2001, pp. 3635-3647
 20. Whitehead, A. and Whitehead, J. **A general approach to the meta-analysis of randomized clinical trials**, Statistics in Medicine, 10, 1991, pp.1665-1677

¹ **Corresponding Author**

cponnuraja@gmail.com

Ph: 91-44-28369632, Fax:91-44-28362525

Postal Address

C.Ponnuraja

Department of Statistics

Tuberculosis Research Centre (ICMR)

Chetpet, Chennai 600031, India

Aura POPA

PhD Candidate, Department of Statistics and Econometrics,
University of Economics, Bucharest, Romania



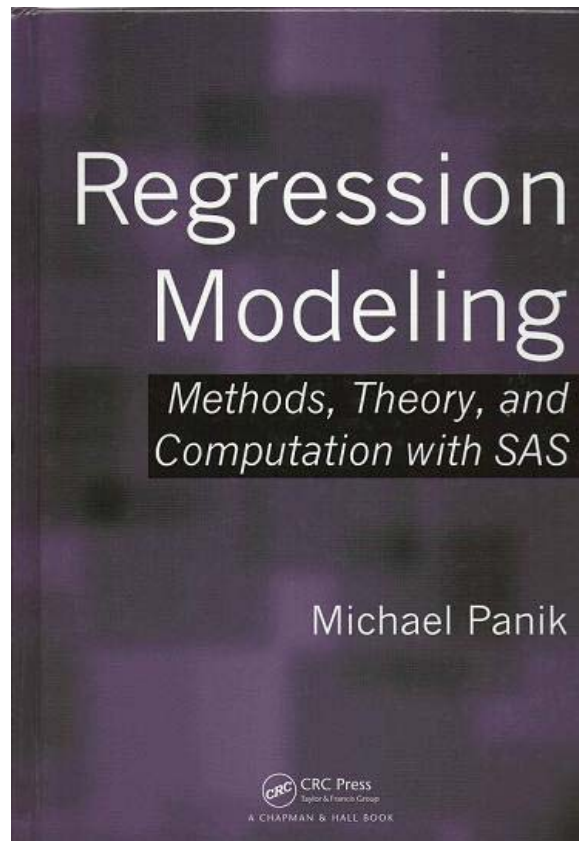
E-mail: aura.popa@csie.ase.ro, **Web page:** <http://www.aurapopa.ase.ro>

Key words: *regression modeling, computation with SAS, multiple regression, fuzzy regression, Michael Panik¹*

**Book Review on
REGRESSION MODELING: METHODS,
THEORY, AND COMPUTATION WITH SAS²,**

**by Michael J. PANIK,
Chapman&Hall/CRC, Taylor&Francis Group,
Boca Raton, FL, USA, 2009**

The First Edition of „Modelling Regression: Methods, Theory, and Computation with SAS” describes both the conventional and less common uses of almost all the regression types in the practical context of today's mathematical, economic and scientific research. This book is designed to introduce the reader to the richness and diversity of regression techniques and is particularly well suited for use in a second course in statistics at the undergraduate or first-year graduate level. This book is a robust resource that offers solid methodology for statistical practitioners and professionals in this field and it is also ideal for students of the applied mathematics or statistics, sciences, economics, and engineering who routinely use regression analysis for decision making and problem solving. Scientists and engineers will find the book to be an excellent choice for reference and self-study. This book blends both theory and application to equip the reader with an understanding of the principles necessary to apply regression model-building techniques in the SAS® environment.



There are a total of nineteen chapters in the book, the first twelve of which form the core, ending in the most well developed chapter, the twelfth one entitled "Multiple Regression". This book has many strengths and important features. It is highly readable, and the material is quite accessible to those enrolled in applied statistics courses or engaged in self-study. In this regard, an objective of this work is to make students aware of the power and density of regression techniques without overwhelming them with calculations. Some common knowledge of matrices and linear algebra is not absolutely essential but, at certain points in the presentation, can be helpful. For convenience, a review of the essentials of random variables, probability distributions, and classical statistical inference is provided in **Chapter 1**. This introductory chapter begins by describing the need for knowing how to apply regression and also reminds the reader the basics of econometric methods.

Illustrating all of the major procedures employed by the contemporary software package SAS, this edition begins with a general introduction to regression modeling, including typical applications. A host of technical tools are outlined, such as basic inference procedures, introductory aspects of model adequacy checking, and polynomial regression models and their variations. The book discusses how transformations and weighted least squares can be used to resolve problems of model inadequacy and also how to deal with influential observations. Although there are many varieties of regression analyses from which to choose, one is most often exposed to ordinary least squares, but this is only a part of the regression story. This text fully exposes OLS and then offers many alternative regression methodologies. Specifically, the regression routines presented here include the following: ordinary least square—along with the method of maximum likelihood, bivariate linear regression and correlation, misspecified disturbance terms, nonparametric regression, logistic regression (including Poisson regression), Bayesian regression, robust regression techniques such as M-estimators, and properties of robust estimators, fuzzy regression, random coefficients regression, L_1 and q -quantile regression, regression in a spatial domain, multiple regression, normal correlation models, ridge regression, indicator variables, polynomial regression, semiparametric regression, nonlinear least squares and some time-series regression issues.

Regression analysis has undoubtedly been one of the most widely used techniques in applied statistics. As a consequence, there are a large number of excellent books written on the topic. And then what makes this book more special than others? Usually authors are bored or avoid mentioning some steps in explaining outputs after running the syntax, but this is not Panik's case! The rigorousness and thoroughness of the evidence in writing this book is appreciated, but for an avid fan of econometrics, many of the chapters can be considered only as a starter in each area, and not a few would expect more, because not only with the `qualitatively` is the reader satisfied, but because of the `quantitatively` that matters.

However, like any book that contains software packages instructions, some mistakes have crept in with the syntax of some programs, and perhaps in the future an errata will be provided to this book that deserves to be cultivated and appreciated, and lead to further reprinting and to also include any further developments because of its high potential.

"Modeling Regression" is well organized. The chapters are sorted in a logical order, from intermediate theory level, computational algorithms, to advanced applications. Some starred sections are advanced and may be skipped for someone who just wants to apply some regression at some point in the SAS Software.

The book is compact, making it easy to extract a feeling of accomplishment and progress and this is making the reader wish to go on, reading more and more, without a large volume of pages being a disarming criteria.

The compactness distinguishes the book from those which try to be too complete and end up being intimidatingly thick. It retains the structure within each chapter. So if the reader is only interested in the example of regression issues related to space, he can jump to the 11th chapter. The presentation of each regression technique is fairly streamlined and designed to offer the reader an unencumbered look into its operation in that proofs and derivations are only supplied in chapter appendices; for those readers who want a more technical treatment, the appendices are a "must read". Moreover, to facilitate the understanding and appreciation of the various regression methods, only the bivariate case is covered in **Chapters 2 through 11**. In these chapters, most of the SAS programs can easily be extended to handle additional explanatory variables once multiple regression is covered. For example, **Chapter 11** provides an introduction to statistical methods for the analysis of spatial data. In a coherent manner, it presents a short classification of spatial data types: geostatistical data, lattice data, and point patterns. Rigorous theory is presented, but not thick, nor boring, but most importantly, bibliographical references are the best, as milestones within each theme separately, and they range from old publications dating from 1935 to date, throughout the entire book and not only in this chapter. Furthermore, regression space syntax can be considered very useful for those who are not good GIS software users, do not use file-type shape, but to reach some notions of space and to analyze the data.

Exercises that are proposed for solution at the end of each chapter are similar to those already solved. For many problems, the data requirements are given which focus more on implementation of syntax, especially running it and interpretation of results afterwards. For most of the regression methods presented, SAS procedure code is included for the convenience of the reader. Although the various sets of SAS code will enable students and practitioners alike to immediately perform their own regression runs, the code given is not all-encompassing, and by no means a substitute for reading the SAS Manuals. It is only intended to give the reader a jumpstart in solving regression problems. Hence, this is not the type of book that only offers theory and proofs; with a modicum of study and effort, one can "hit the ground running", so to speak, and readily generate some fairly sophisticated regression results. Once a regression technique is explained, SAS handles the "how to" portion of the presentation. This is imperative because one can first study a regression method and then, for the most part, directly apply it because numerous example problems are included, with the SAS results explained in considerable detail.

Also, plotting is great, with ample illustrations, explained, closely connected to the theoretical and practical part. Even though it is printed in black and white, the graphs are so well explained and exposed, that you do not miss the lack of colors in them.

In particular, I appreciated some topics that are covered when we are talking about regression. One involves some highlights over spatial domain. The author might recommend this chapter for further research, because how it was presented: in a short fugitive way in comparison with other chapters, and because this area can not be exposed, even punctual in approximately ten pages. The other involves the excellent description for details of fuzzy theory and how to implement this large area into computer area.

I would appreciate it even more if this book would have had a dedicated website. Being published only at the beginning of last year, the publisher could invest more time and

could create a virtual area for the author to give the opportunity to provide for his readers access to several datasets, software, and probably other useful materials, sparing some extra paper, pages and reader's money.

In summary, I can recommend this book as a very good graduate level text-book. I also believe that readers will learn much more not just reading it, but using it with their computer, applying each exercise that the author is proposing.

¹ Michael J. Panik received his PhD in economics from Boston College and, as a doctoral student, he held a NASA Fellowship for three years developing specializations in the areas of statistics, econometrics, mathematical economics and microeconomics. He also held a lectureship at Boston College and then taught for many years in the Department of Economics at the University of Hartford, where he is now Professor Emeritus. Dr. Panik has been a consultant in the area of health care research and to the state of Connecticut, and has written numerous articles in professional journals and four other books in the fields of linear programming, convex analysis, optimization and statistics. His current research involves themes like growth curve modeling, estimation, and also analytical microeconomics.

² **Acknowledgements**

This book presents many SAS programs which are useful only if you are able to access this software, otherwise the reader has to restrict himself only to the theory section which is in strong connection with the application part. That is why special thanks are due to the SAS Centre of Excellence from the University of Economics-Bucharest, Romania for making it possible to access this precious software without which this book couldn't be properly evaluated, and that with their help, quantified in technical support, they pushed the Romanian academic research and teaching one step further ahead.

Cristian CIUREA¹

PhD Candidate, Department of Computer Science in Economics,
University of Economics, Bucharest, Romania

E-mail: cristian.ciurea@ie.ase.ro



Key words: RIA; Flex; ActionScript; C#

**Book Review on
DEVELOPING RIA APPLICATIONS
("DEZVOLTAREA APLICATIILOR RIA"),**

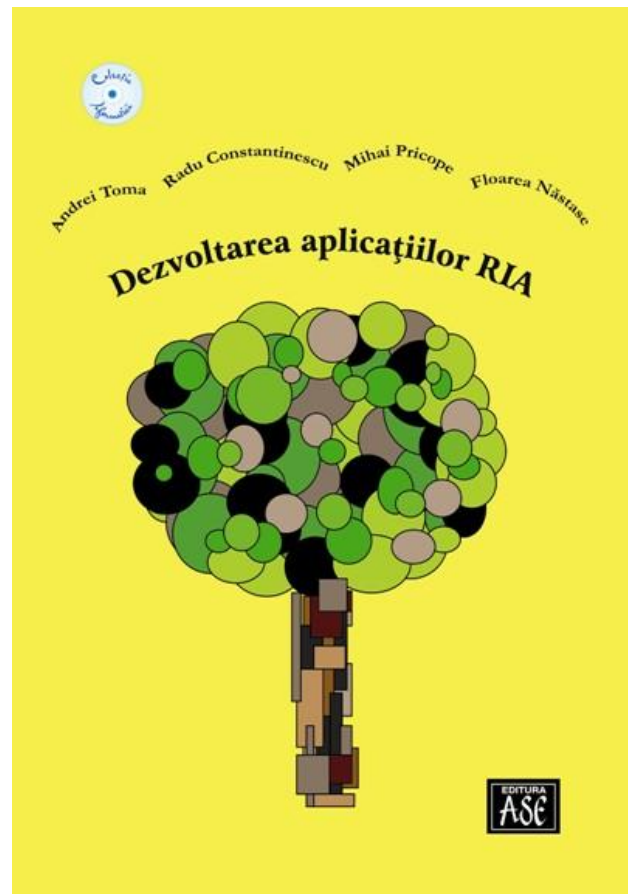
**by Andrei TOMA, Radu CONSTANTINESCU,
Mihai PRICOPE, Floarea NASTASE,
ASE Publishing House, Bucharest, 2010**

The book *Developing RIA applications*, has been published in 2010 at ASE Publishing House, Bucharest, and is a book with a specialized language. It is addressed to Web application developers, both those experienced and those who want to learn Adobe Flex technology.

The book is useful for web application developers with knowledge of C# and addresses the problem of building applications with a rich interface based on data access services.

Rich Internet Applications allow not only optimizing current processes and refining the appearance of web applications, but also using mechanisms that were typical to desktop applications.

In order to illustrate the concepts needed to develop a Flex application, an online auction site is built step by step in the book. The need for this approach comes from the ease of use benefits of RIA client applications compared with the clients built on a classic request-response paradigm. The client application is developed in Adobe Flex, a mature technology for the development of RIA applications and the server application in ASP.NET and C# with data stored on MS SQL server.



The book is structured in 12 chapters, very well divided and having a logical sequence. The book deals with issues related to building and customizing a Flex interface oriented on ease of use, data access through Web services, defining custom components, access to external services etc. On the server component development, building Web services and data access in C# are addressed, with problems related to construction and processing of XML results.

The first chapter is dedicated to the presentation of the technology necessary for web application development. Both client-side technologies, such as HTML, XHTML, CSS, JavaScript, as well as server side, such as PHP, ASP.NET, Python, JSP and ColdFusion, are discussed.

Chapter 2 makes an introduction to Adobe Flex, understanding RIA concepts, Flash Player, MXML and ActionScript.

In Chapter 3 Flex interfaces are presented, Adobe Flex being a technology that implements model-view type architecture.

Chapter 4 is an introduction in the ActionScript and MXML. The MXML language is the Flex version of a specialized language for interface construction. ActionScript is a language that defines data types strictly and statically, supporting a range of data types, some primitive, others complex.

In the context of the 5th chapter, web services in C# are presented, the reader is shown how to create a web service, followed by data access and data update through C#.

Chapter 6 describes data services in C#, theoretical concepts needed to extend the Web service to retrieve data, construct and return an XML structure of categories.

Chapter 7 is dedicated to Web services and data sources in Flex, also covering the topics of events and Data Binding. The chapter covers data access through Web Services and implementing data access related events.

Chapter 8 contains web services, data sources in Flex, and explains how to upload images on server.

Chapter 9 presents the concepts of Data validation, Item editors, Item renderers, Drag and Drop.

Chapter 10 covers working with Sessions, Skins, Effects and Transitions, Custom components.

In chapter 11 external APIs and Mash-ups are shown, understanding SOA concepts and external services and how to use Yahoo mapping service.

The last chapter deals with architectures, the problem of software applications design and security of Flex applications.

The topics addressed in this book cover the introduction to building RIA applications with data services; with the information contained in the chapters the construction of a complete application is possible. Besides the subjects related to the effective development of medium complexity applications, issues related to refining an application are dealt with (such as reducing project complexity and handling security concerns).

¹ Cristian CIUREA has a background in computer science and is interested in collaborative systems related issues. He has graduated the Faculty of Economic Cybernetics, Statistics and Informatics from the Bucharest Academy of Economic Studies in 2007. He is currently conducting doctoral research in Economic Informatics at the Academy of Economic Studies. Other fields of interest include software metrics, data structures, object oriented programming in C++ and windows applications programming in C#.