WWW.JAQM.RO

JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

Vol. 11
No. 3
Fall
2016

ISSN 1842–4562

JAQM Editorial Board

JAQM Advisory Board

Contents

Page

Quantitative Methods Inquires

BIG DATA: ISSUES AND AN OVERVIEW IN SOME STRATEGIC SECTORS

Massimiliano GIACALONE[1]
PhD, Researcher, Department of Economics and Statistics,
University of Naples Federico II, Italy

E-mail: massimiliano.giacalone@unina.it

Sergio SCIPPACERCOLA[2]
Associate Professor, Department of Economics, Management, Institutions,
University of Naples Federico II, Italy

E-mail: sergio.scippacercola@unina.it

Abstract

Big Data is a new technology with a model that works with a large amount of various type data (structured, semi-structured and unstructured) differently from static data being stored in warehouse. The data are generated from a variety of instruments, sensors and mainly by computer transactions. They are constantly updated with a high frequency and become more and more accurate and precise with the passage of time. Main purpose of this paper is to bring into light the new technologies, process and statistical analysis to extract values and results from Big Data. This work, in the first part, introduces the main characteristics of Big Data and its basic management. Important suggestions are developed for a quality control before to extract significant samples for subsequent analysis. Follows, in the second part, the comparison with other traditional techniques. In the last part, the paper highlights the growing role of Big Data and the key benefits in some strategic sectors (Education, Health Care and Banking Industry). Common to our interest fields, the principles of ethics and privacy, to be observed, are also mentioned.

Key words: Big Data, Data Quality, Data Mining, E-learning, Learning Analytics, Health Care, Banking Industry, Ethics, Privacy

Introduction

Big Data were born because of the massive proliferation of elementary data from multiple sources. Sets of images, e-mail, GPS data, or information obtained from web sites (such as access, permanence, etc.) can be defined Big Data (Snijders, et. al., 2012). One of the fundamental characteristics of Big Data is the heterogeneity of data sources: these are data sets, or frequently of dynamic flows of 'metadata', from heterogeneous databases (Rezzani, 2013). Table 1 shows, in a non-exhaustive list, the main sources from which they are taken Big Data. Almost all data is generated typically at sub-daily basis: hours / minutes / seconds / milliseconds. For example, not only censuses, surveys, interviews, or questionnaires; but also information collected from the Internet, by telephone networks, by satellite, or for transport, may be part of the same set of data.

Big Data is often confused with simple digital traces of human activities mediated by information and communication technology, as are the recordings of access to services which are called the log of service (phone calls, messages exchanged with identification of the applicant, short texts, and geolocation). Even the logs can be considered data of a Big Data system.

Table 1. Some Data Source of Big Data and type of data

Data source	Type of data generated
E-mail, SMS, instant message, YouTube, WhatsApp, Web	Textual, graphical, and audio video
Electronic medical instruments, scientific experimental and observational data	Numerical (i.e. temperature, pressure, etc.), and diagnostic images (i.e. computerized tomography, ECG, etc.)
Environmental sensors	Numerical, textual, graphical, audio-video
Financial transactions	Textual and Numerical
Traditional Database and Datawarehouse	Numerical and textual
Satellite	Numerical and graphical

A common definition of Big Data is that offered by Doug Laney (Laney, 2001), which is based on the paradigm of the three V (Volume, Velocity, Variety) (Table 2):

- **Volume**: it is estimated that by 2020 a measure of 35 thousand billion gigabytes of data will been generated. With regards the enumerations of the bases of Big Data, it the measure would have gradually proceeded to extend the measuring units that gradually proceeded in extension of the average-sized volumes in place, arriving today with volume orders of magnitude expressed in 'Zettabyte', equal to one billion terabytes and Yottabyte equal to one trillion of bytes.
- **Velocity:** Once extracted, the data must be analyzed promptly, not to become obsolete, and therefore unnecessary to make a "decision". The fast acquisition and access to required data is therefore essential. Just think that it is not uncommon the need to acquire 'live data' (for example, access to sites, search engines in the Internet, or share data in television), to process on a daily basis, and, mainly at sub-daily basis.
- **Variety:** Data have highly heterogeneous nature (eg., texts, images, videos, web searches, financial transactions, email, post on blogs and social networks, etc.), and each size requires a dedicated treatment. This characteristic of Big Data may require scaling operations or conventional classifications (for example catalog of images for the chronological date, or for chromatic scale or according to another ordinative scale) (Manyika et al., 2011).

Some scholars suggest adding to the definition of Big given two more V:

- **Variability**: the data must be contextualized, as their meaning can vary depending on the context.
- **Virality:** the growth of Big Data is exponential, like wildfire.

These peculiar features and specifications require that, with respect to storage, the constituents Big Data, are both structured to unstructured, and are expressed on different measurement scales, or are also qualitative (Table 2).

Therefore, Big Data is not just a lot of data, but it is a **System to handle a large amounts of data of any type.**

Table 2. Features of Volume, Velocity and Variety of Big Data

Feature	Size, Time and Type of data
VOLUME	Size: TB (terabyte = 1024^4 byte) - PB (petabyte =1024^5 byte) - EB (exabyte = 1024^6 byte) - ZB (zettabyte =1024^7 byte) -YB (yottabyte = 1024^8 byte)
VELOCITY	Time: Results in real time: fast acquisition and access to data is essential, especially for 'live data' that must be processed on daily or sub-daily basis.
VARIETY	Type of data: Structured – Semi-structured – Unstructured (qualitative)

In this complex panorama, the aim of this work is to bring into light the new technologies, process and statistical analysis to extract values and results from Big Data.

The paper is structured into three parts: introduction to Big Data, comparison with other traditional techniques, main applications to some strategic sectors.

In the first part, after the present introduction, the difficulties inside the Big Data management are presented. Due to the large amount of data produced in continuity and to the need to work on samples drawn from the population it is essential to carry out a preliminary Data Quality Statistical Control (Section 2).

In the second part, we compare the Data Mining methods already used for some time, with the new frontiers opened by Big Data (Section 3). This part is devoted to the applications of Big Data in strategic sectors as E-learning, Learning analytics, Healt Care, Banking Industry (Sections from 4 to 7).

In the last part, the principles of ethics and privacy in the era of Big Data are discussed (Section 8) and the main benefits of using Big Data in the analysed sectors are underlined (Section 9).

Finally, in the conclusions, the perspectives that today offer the Data Science including Big Data are emphasized.

2. Quality Assessment Process for Big Data

The management of Big Data is very complex because many are the data stored and sometimes the Big Data are erroneously also referred to as large the data set. We must filter from this very lot of data selecting only those that meet the quality control requirements. The filtered data become statistical sample to which it is possible to apply inference or traditional analysis methods of Data Mining. Otherwise, for work directly with Big Data we need only apply special parallel algorithms. In this context it is useful, also, the adoption of 'genetic algorithms' capable of operating a meeting of non-metric also data from dynamic sources coagents in the process of dataset formation. These algorithms can be used for categorical or probabilistic selection methods (selection of 'Boltzmann') (Koza, 1992; Wright, 1991).

The progress made in the meantime by the scientific and technological research in hardware and software area have ensured satisfactory performance in terms of efficiency, access to Big Data and power and effective processing speed.

Big Data collection requires to acquire and analyze data from several sources and with various researchers. For this reason decision-makers have gradually realized that this massive amount of information has benefits for understanding customer needs, improving service quality, and predicting and preventing risks. It is also logical that use and analysis of accurate high-quality data is a necessary condition for generating value from Big Data. Therefore, we analyzed the challenges faced by Big Data and a quality assessment framework and assessment process for it.

In the last years, Xi'an Jiaotong University set up a research group of information quality that analyzed the challenges and importance of assuring the quality of Big Data and response measures in the aspects of process, technology, and management (Zong & Wu, 2013).

Big Data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn. This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it.

An appropriate quality assessment method for Big Data is necessary to draw valid conclusions. In this paragraph, we propose an effective data quality assessment process with a dynamic feedback mechanism based on Big Data's own characteristics, shown in Figure 1.

Different tasks like filtering, cleaning, pruning, conforming, matching, joining, and diagnosing should be applied at the earliest touch points possible.

After the quality assessment preparation is completed, the process enters the data acquisition phase. If the analysis results meet the goal, then the results are outputted and fed back to the quality assessment system so as to provide better support for the next round of assessment. If results do not reach the goal, the data quality assessment baseline may not be reasonable, and we need to adjust it in a timely fashion in order to obtain results in line with our goals. Poor Big Data quality will lead to low data utilization efficiency and even bring serious decision-making mistakes.

The application of SPC methods to Big Data is similar in many ways to the application of SPC methods to regular data. However, many of the challenges inherent to properly studying and framing a problem can be more difficult in the presence of massive amounts of data.

There exist several frameworks for solving problems in the Total Quality Management (TQM), Statistical Process Control (SPC), or Six Sigma area (Montgomery, 2013).

The classical tools are the Plan Do Check Act (PDCA) Deming cycle, or the Define, Measure, Analyze, Improve, Control (DMAIC) cycle can be applied to Big Data (Qiu, 2014). For example, the Cross Industry Standard Process for Data Mining (CRISP-DM) and knowledge discovery in data mining (KDD) were recently introduced.

It is important for researchers in statistical surveillance to consider processing speed when developing and refining methodologies. Another challenge in monitoring high dimensional data sets is the fact that not all of the monitored variables are likely to shift at the same time; thus, some method is necessary to identify the process variables that have changed (Megahed & Jones-Farmer, 2015).

Another important challenge when using SPC methods with big data applications is that, traditionally, SPC methods were developed for numeric data. While there are some attributes control charts, these tend to be a distant choice to using methods designed for quantitative variables.

However, one of the great challenges of big data is the ability to process and analyze unstructured data. Most of big data applications are concerned with non-numeric data obtained from several databases.

In the next future, more complex hierarchical structure of a data quality system could be analyzed and proposed to evaluate the Big Data quality framework.

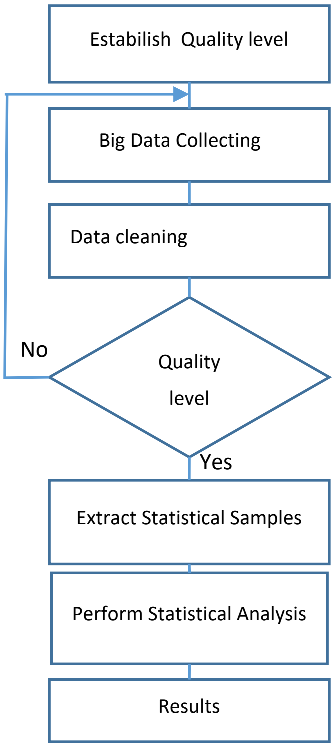


Figure 1. Big Data Statistical Quality Control

3. Data Mining and Big Data

Data Mining is part of Business Intelligence, and indicates the process of exploration and analysis of a set of data to identify any regularity, extracting new knowledge and meaningful applicantion rules.

The main objective of the "Data Mining" is to "extract information" useful from a database and turn them into a data structure (pattern) for further use survey. Among the main applications of Data Mining we can highlight the summary description of the data, the associations and correlations, classifications, and evolutionary analysis (regularity of data that changes over time). The techniques of data mining are adopted in various fields as Statistics, Sciences of Education, Economics, Medicine, etc.

There are clear similarities found among the "Big Data" and "Data Mining". The latter could be considered the **old Big Data** because it responds at least in part to two of the characteristics of Big Data that are the size and velocity, but lacks the third V (variety) as the Data Mining is often extracted knowledge only by means of the Database or Data Warehouse and Data Mart that are retrospective static type unlike Big Data constantly updated with a high frequency and become more and more accurate and precise with the passage of time. Therefore, Data Mining could be considered the **old Big Data** and Big Data could be considered the **new Data Mining.**

Another interesting aspect of Big Data that differentiates it from the Data Mining is the **structural diversity** (Fig. 2). Some data have a well-defined format, in the classic way of files / records / fields, such as, for example, in the transactions recorded in a database other data may be of very different type (i.e. Municipal data, Driving Licence Data, etc.); such as a collection of texts on a blog, or tables, or images, or audio recording, or video.

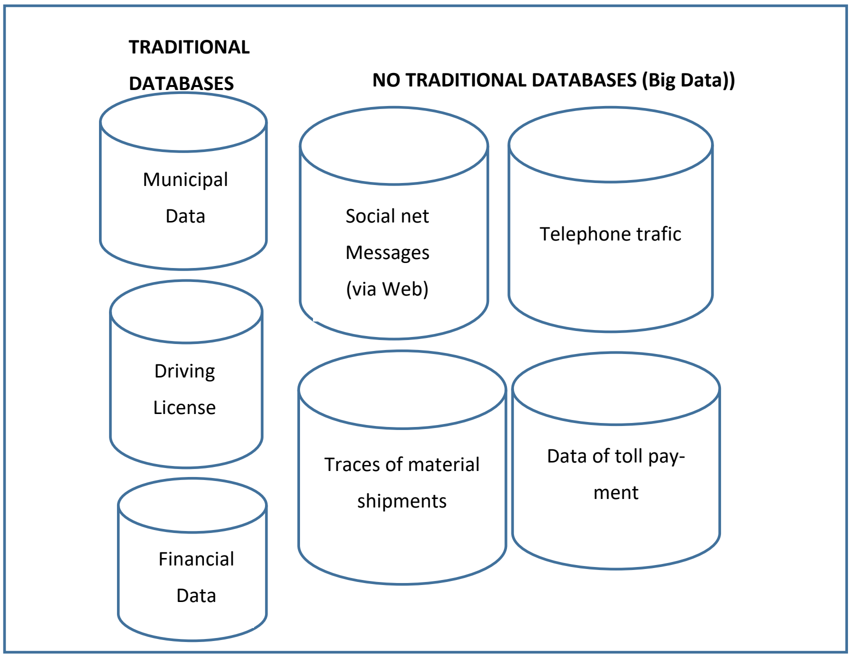


Figure 2. Examples of structural diversity between Data Base and Big Data

From the point of view of the architecture and engineering of the dataset and data structures, the latest models of Big Data are based on highly scalable methodologies, and type of **No Structured Query Language** (***NoSQL***) solutions (Vaish, 2013). It is intended for *No Structured Query Language* a set of technologies forming a different new data management system from the traditional *Relational Data Base Management System*, because the relational model is not used, it does not have an explicit scheme and the system is designed to work quickly and well in the cluster.

The Big Data have redundant informations (**redundancy conditions**) and it is preferable to work with samples. A preliminary inferential approach to aggregation comes beore the actual creation of databases and datasets useful for the statistical analysis (Manoochehri, 2013).

Briefly, the aggregation of mixed numerical sources is addressed by operating on data streams in parallel (*approach map*) then subjected to reduction treatment, filtering and 'clean' data eliminating those untrue or unnecessary (*Data garbage*) before to operate combinations and reorganizations in the final dataset (Reiss et al., 2012).

4. Big Data for e-Learning

The impact that Big Data in education - both with reference to teaching, which learning - is relevant, not only in the design of the modules, but also in terms of refinement of learning objectives already predefined (Gutierrez-Santoz et al., 2012). The Big Data can be used in multiple sectors and e-learning is one of them.

The traditional or *e-learning training* can be evaluated at four progressive levels (Kirkpatrick, 1979) (see Table 3, with our adaptation to domain of the education). Before the delivery of the training, we should have identified our strategy, completed an assessment and built a plan. Then, during the delivery of the education solution, we need to manage a number of factors to ensure success. After the delivery, we have to evaluate the success of the implementation in terms of the originating need and strategy (Giacalone, Scippacercola, 2016)

The E-learning teaching materials should be built ad hoc to ensure the four main characteristics of online education:

- Modularity: course material should consist of "learning modules", also called *Learning Objects;*
- Interactivity: the student must interact with the system by providing his answers that are properly recorded;
- Exhaustiveness: each module should contain a complete topic;
- Interoperability: instructional on any platform and technology to ensure traceability of the training.

Currently the most common standard is the Shareable Content Object Reference Model (SCORM) (Bohl et. al, 2002). Technological progress has led to the creation of the Learning Content Management Systems (LCMS) that deal with the content management both in the process of creation and during the delivery: they can be considered a complete platform for e-learning. Today we are able to track and collect this data also through social networks and any other media.

Table 3. Evaluation of e-learning training

Action	Evaluation and measurement of
Reaction	personal reaction to the training
Learning	the increase in knowledge
Behavior	changes in on-the-classroom behavior
Results	obtained vs desidered results

Each time that the learners (students) interact with the content of a course, in fact, they produce data.

Beside the usual 'assessment of end-over', by means of the satisfaction questionnaires proposed to learners, it grows and becomes relevant the need to acquire real-time information always more detailed and organized on the various areas of teaching evaluation. For example, accesses ('visits') to Web pages are data that can be purchased on-line with other data, to compose patterns useful for teaching evaluation.

By Big Data, the e-Learning responsible teachers can receive information to make teaching more effective, or to correct any defects. For example, access to websites, the data collected from social networks, the content of the web searches, and online learning modules, Big Data can be useful to assess the information use by learners and their behaviors in the learning phase.

An interesting prerogative is given by the possibility, using special software programs or power tool immediately discard the data not useful from the information point of view. The use of mathematical models and statistical methods on data of e-Learning, once organized the same in databases or 'metadata', allows us to produce models of understanding or even useful prediction refinement or simple evaluation of teaching methods (Chatti et al. 2012).

Another approach to the use of Big Data, is to evaluate different parameters of prefixing didactic training for each variable appropriate 'threshold values' or 'levels-target' to achieve the educational objectives. (Siemens et al. 2011).

Christopher Pappas (Pappas, 2014) listed in this regard five benefits that can be drawn from the analysis of data related to the use of a course and e-learning:

1. The data analysis allows you to identify which type of teaching is more effective in achieving the objectives of the course.
2. It becomes possible to identify improvements of the educational path. For example, if a large number of learners taking too long to complete a certain module, means that the form must be made more slender and usable
3. And it is possible to monitor what are the forms displayed the most shared links with other learners.
4. The data resulting from the traces of the learner are immediately available and there is no need to wait for the final evaluation of the test results to know the situation. In this way, teachers can get an overall picture of learners' behavior and can optimize the learning strategy in near real time.
5. Based on the data it is possible to make predictions about the successes and failures of learners and develop in a way that courses that students have always the possibility of obtaining the best possible result (Pappas, 2014).

In summary, the main advantage of collecting and analyzing Big Data in e-learning, is mainly the possibility of obtaining useful information to customize the learning experience based on the needs and learning styles of learners (Giacalone, Scippacercola, 2016).

5. Big Data for learning analytics

The term *learning analytics* identifies an important sector within the Technology-Enhanced Learning emerged in recent years and is closely related to several disciplines such as Business Intelligence, Web analytics and Educational Data Mining (EDM). The term learning analytics refers to the measurement, collection, analysis and presentation of data on students and their contexts for understanding and optimization of learning and the environments in which it takes place (Baker et al., 2014) (Ferguson, 2012, 2014).

The transfer of the knowledge through *learning objects* in the various environments is missing of reference standards for the assessment. There are various products (like Google Analytics, Omniture SiteCatalyst, WebTrends, Coremetrics, etc.) that allow the retrieval of information transmitted over the web.

The evaluation of the dissemination of knowledge via web can be done by traffic parameters that can be listed as inferred from such a traffic controller that can detect what in slang is called the *Visitors Overview*. This traffic overview allows you to view in detail the aspects of quality (ie average pageviews, time spent on site, bounce rate) and characteristics (for example, first time visitors, return visits) of visits. The traffic indicators can be classified in two types (Scippacercola, 2012):

– indirect (the number of accesses to the module, the usage time of a session, the mode of use, flow of the navigation in the website, etc.)

– direct (the average response time to questions, the number of attempts before you answer correctly, etc.) The direct indicators are derived often by user surveys.

Exist metrics that allow evaluating, ex-post, the personal reactions to the training and permit to evaluate the validity of the a web page or for directing the eventual reengineering. Assume that a web site (with four pages P_i) (i = 1, 2, 3, 4), illustrated schematically in Fig. 3 in the box inside, is visited by three (A, B, C) hypothetical students that enter the website, and consider, for example, the following actions:

- Student A sees P_1, P_2, P_1 and then exits from the website;
- Student B sees P_4, P_2 and then exits;
- Student C sees P_3 and immediately exits.

Using the above indicators it is possible evaluate the traffic of students and the analytic behavior on the same network.

Referring mainly to the most widely used Web analytics (Google) (Clifton, 2010; Vasta, 2009) we list the main indicators that it gives us (Fig. 3) (Scippacercola, 2012):

- *Entrance*: the *number of inputs* to the page P_i**;**
- *Pageviews*: is the *total number of requests* for loading a P_i of the website;
- *Unique Pageviews*: is the *number of sessions* in which a page was viewed more than once;
- *Average Time on Page*: is one way of measuring visit quality. A high Average Pageviews number suggests that visitors interact extensively with the web site;
- *Bounce Rate:* is the *percentage* of single-page visits (i.e. visits in which the person left your site from the entrance page). The percentage of visits where the visitor enters and exits at the same page without visiting any other pages on the site in between. The Bounce rate is one way of measuring visit quality. A high bounce rate generally indicates that the entry pages (landing) is not relevant to your visitors.
- *Exit*: is the *percentage of users* who exit from a page.

The approach here considered can be classified as a theoretical approach to **ex-post non-interactive**. Conversely other approaches tend to interact during the learning phase (**interactive approach**). In Ferguson is reported, for example, the Signals Project (Ferguson, 2014), developed by the Purdue University explores large datasets and apply statistical tests to predict, during the courses, students who risk being left behind.

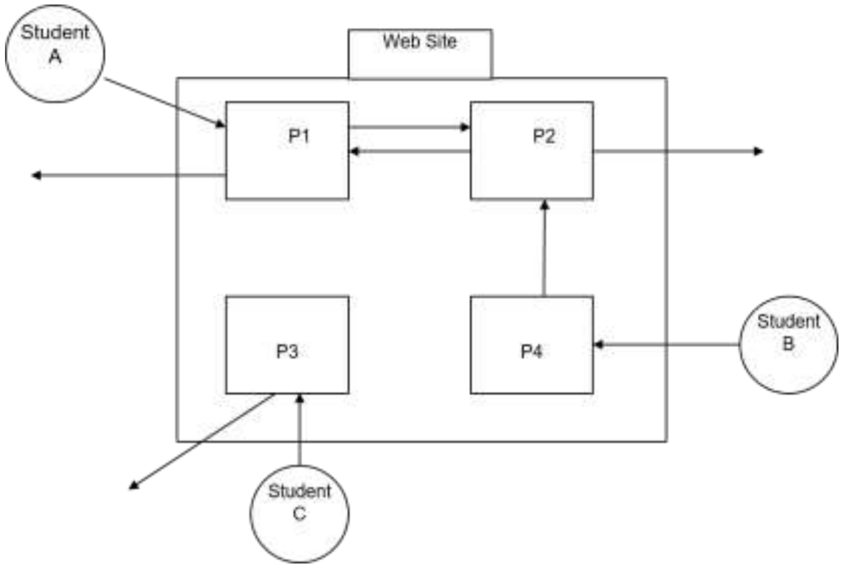


Figure 3. Students which link to a website to make navigation in four pages (P$_1$, P$_2$, P$_3$ and P$_4$)

The goal is to produce actionable intelligence, guiding students to appropriate resources and explaining to them how to use them. A traffic light shows students if things are going well (green), or if they were classified as high risk (red) or moderate risk (yellow) (Alan et al., 2010). The reported results are promising although the system as data and software may not be entirely comparable with a Big Data system.

From a technological point of view learning analytics is an emerging discipline and its connections with Big Data, despite some significant proposals in American College (Picciano, 2012), it remains to be developed.

6. Big Data in Health Care System

In the Health Care sector the information is the most important aspect, and the human body, in particular, is the major source of production of data. Consequently, the new challenge for health care world is, knowing how to take advantage of these huge amounts of unstructured data between them. Electronic medical records include within them data with each other very heterogeneous in terms of size: audio recordings, magnetic resonance imaging, computerized tomography and other diagnostic images, electrocardiograms, and the list goes on indefinitely. Nevertheless, electronic medical records have to be designed to process and manage data characterized by high volumes, generating speed and wide variety of sources (Sanchez et al., 2014) (Murdoch et al, 2013).

Organize Big Data health means being able to sort the huge amount of information about the medical history of each patient. A concrete example is the electronic medical file that will soon replace the medical records. A single support will allow the patient to store in one device prescriptions, drugs, diagnostic tests, laboratory analysis findings, emergency department, hospital, and the doctor to rebuild quickly and accurately the state of overall health and especially the patient, in addition to being able to share information with other doctors in the case of diseases that require more expertise.

The *Big Data analytics*, cloud computing, social networking and the emergence of micro-sensors are the main technologies improving predictive analysis in the medical field and the patient's quality of care. The traditional Data Warehouse strategies are not easily and quickly scalable and they provide a retrospective view and not in real time or predictive. Through data analytics we can classify data, make predictions, and greatly increase the understanding of patient's clinical data (Raghupathi, 2014).

The **Clinical Intelligence** (Fig. 4) (Groves et al., 2013) consist of all the analytical methods, made possible through the use of computer tools in the set of processes and disciplines of the mining and processing of raw clinical data into meaningful insights, new discoveries and knowledge that ensure greater efficiency clinical and better health-related decisions (Harrington, 2011). Clinical intelligence is the set of electronic methods, processes and disciplines extraction and transformation of raw data into meaningful clinical insights, new discoveries and knowledge that affect the clinical decision-making and the decisions in the health sector.

The clinical intelligence differs from *business intelligence* for the following considerations. The business intelligence deals with raw economic data, often structured, and provide insights and information on the decision-making process in the economic field. In contrast the clinical intelligence deals with clinical data and requires statistical methods and analysis much more sophisticated than that used by business intelligence.

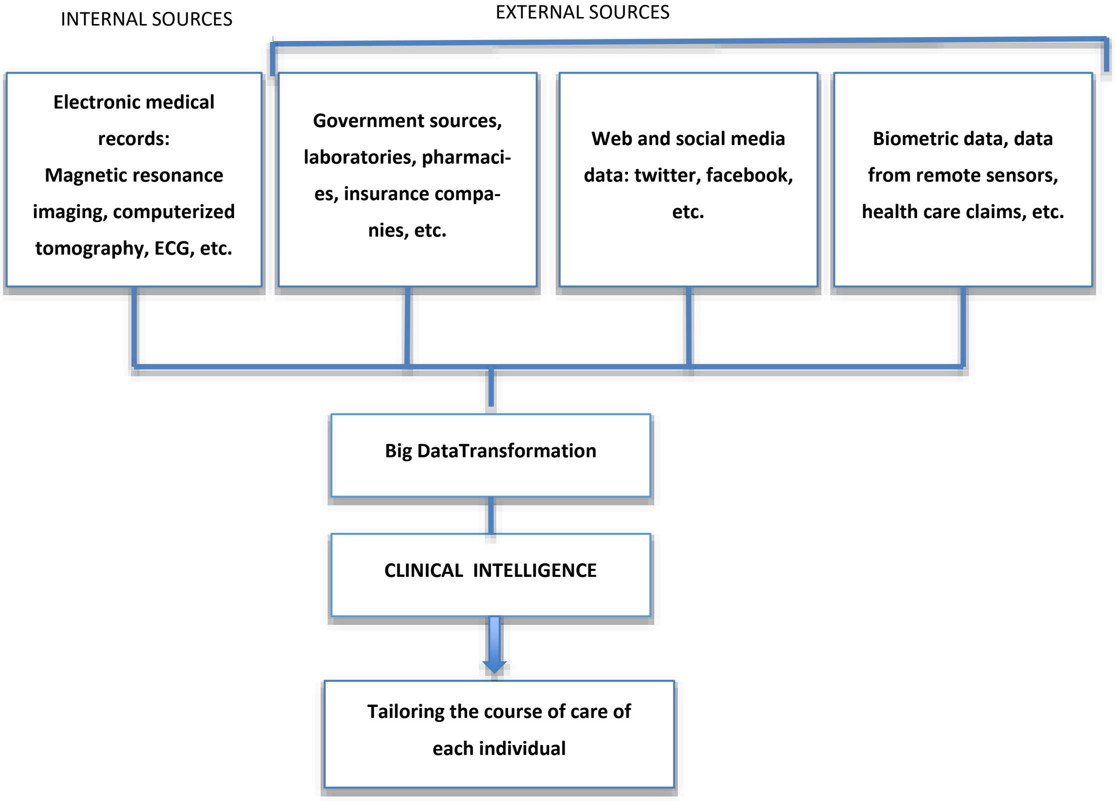

INTERNAL SOURCES
EXTERNAL SOURCES
Electronic medical records: Magnetic resonance imaging, computerized tomography, ECG, etc.
Government sources, laboratories, pharmacies, insurance companies, etc.
Web and social media data: twitter, facebook, etc.
Biometric data, data from remote sensors, health care claims, etc.
Big DataTransformation
CLINICAL INTELLIGENCE
Tailoring the course of care of each individual

Figure 4. Some main sources for Clinical Intelligence

The clinical intelligence often has to do with types of data unstructured and much more complex as data that often arise ambiguous, incomplete, conditional and inconclusive. The clinical intelligence uses sophisticated methods to study the data and interpret the results as machine learning techniques, non-linear and multi-algorithm approaches. The clinical intelligence allows a more sophisticated classification of patients' based solely on demographic variables such as age, sex, lifestyle but also on relevant medical and clinical features related to certain diseases, medical conditions, genetic predispositions and the likelihood of therapeutic response (Chawla et al., 2013).

The clinical intelligence makes it possible to optimize and custom *tailor the course of care of each individual* patient by basing it on a multitude of factors that define the medical protocol of care: previous medical history, known allergies, personal risk factors, genetic traits, lifestyle and business, management of personal safety. The clinical intelligence allows implemention of multi factorial analysis to determine the effective utility associated to different treatment courses (Kayyali et al, 2013). These analysis allow doctors to identify the most appropriate treatment for a particular patient as well as specific indicators for measuring outcomes. Analytical tools, when applied to medicine, are able to suggest medical care plans and clinical pathways providing a prediction of the results corresponding to them. These tools allow you to compare treatment options which are complementary, manage the risks associated with each treatment plan and select the most appropriate course of care (Murdock et al., 2013).

Therefore, Big Data are radically changing the world of health and medicine (Raghupathi et al., 2014; Growes et al., 2013), allowing for more personalized care pathways (patient experience), effective and less subject to clinical risk through new mechanisms for control and governance of processes. In the opposite, while the structured data provide the "what" of a disease or medical treatment, rarely they offer a "why" behind the decisions taken. In many cases, the collection of unstructured data remains the best option to capture the details in depth, for example, a medical record as they contain valuable information on the health of the patient.

7. Big Data for banking

Banks generate a large amount of data: paper documents with signatures, checking accounts, mobile banking, credit and debit cards, loans, etc. Today many data comes from the Bank's contacts with customers in *online mode*. Such data can be classified into structured data like e-mails, chat logs, feeds, posts, web logs and semi-structured data such as customer reviews. Among others the bank's objectives, we present, in particular, the **fraud detection** (Russom, 2011; Hipgrave, 2013), the **money laundering** and the **risk analysis** (Boinepelli, 2015).

Some of **frauds in the banking** are in the "online banking (credit card, internet or mobile transactions)" where the fraudster performs the transactions with the same code or sign of customer. The bank, in order to prevent the success of such fraudulent transactions while running you must provide a customer profile based on the story of its financial transactions. Through the analysis of all the transactions made by customers you can create a "**customer profile**" and its relations with other correspondents and payment methods used. Each new transaction is compared to the profile to identify if an operation does fit into the usual ones and if not, to report suspicion of fraud. To create the profiles are indispensable and techniques ranging from simple statistics (mean, standard deviation, minimum and maximum values of the transactions, dates, etc.) to advanced (inference) will be required.

Money laundering continues to be a local, regional and global concern. To combat money laundering is essential to have databases containing multidimensional data on financial transactions and the database of the police. Clustering, classification, outlier identification, data visualisation tools are used to detect behavior fraudster in transactions with large amounts of money between accounts. Both of these databases are the platform to reveal associations and patterns of activities that help identify the fraudster suspicious.

The fraud detection is in real time by means of suitable software that should be used for example to prevent unlawful claims before they are completed, and analyze business data where fraud has been committed in the past.

The third application of Big Data in the banking is the **risk analysis**. The predicton of default on loan and credit card accounts is important for banking. The model that allows you to make predictions includes a data warehouse that contains historical data of customers, the realized contracts and the observed results. The model provides for the application of traditional Data Mining techniques so as to achieve a financial risk analytics framework. In addition to the Data Warehouse in the model are considered all contacts between the customer and the bank in the period between the loans granted and the end result. Using the results obtained from the aforementioned model, the bank can identify and classify more safely customers with a risk of insolvency.

8. Ethics and privacy

The Big Data collection activities have different effects in the sensitive area of the issues pertaining to the processing of personal data, called 'sensitive' under the legislation and the overall system of constraints and responsibilities, governed by privacy laws.

Relevant are the ethical issues arising from the management of personal data of users, with reference also to the possibilities of diffusion, sharing and use of information 'sensitive' about learners themselves.

There are critical issues that have already led professional organizations to a serious reflection on the contents and on the delimitation of the operational boundaries of e-learning activities in order to reach the fulfillment of plans the collection and interventions 'legal' ie capable of qualifying the various aspects of the activities in place (the collection, conservation, management, use and publication of personal data) in observance of 'best practices' default (Slade et al., 2013).

One of the most important issues is the communication of the intent and purpose of the collection of data for the sectors analysed, in order to achieve preliminary authorization and legitimacy in accordance with local legislation, making clear communication to users interested in each of the which require explicit consent to the processing of their personal data.

If there are 'stakeholder' or external customers of data collection, in the same way the organizations and professionals that carry out studies or data analyzes in the specific sector should carry out all acts of communication and producing relative contracts in accordance with applicable regulations in order the processing of personal data, both with regard to any estate regarding the dissemination of personal data obtained or evicted from the activities in question.

Another aspect is not secondary to the safe preservation and accessibility of personal data in a server equipped with procedures, protocols, and active and passive safety standards, as required by the regulations in force for years, and international safety standards (ISO, EN) (Corposanto et al., 2014).

With regard to the preservation and accessibility of data, technologies in support of Big Data are highly reliable, low-cost and scalable. For example, Hadoop (Hadoop, 2014) is an adequate system to Big Data, because it allows you to store huge volumes of data and then process them when more is appropriate. Hadoop also allows the distribution of data on multiple nodes, reducing the computational and storage costs for storage and analysis of Big Data, and masking hardware failures. It has been estimated by Zedlewski (Zedlewski et al., 2003) that the cost of a data management system based on Hadoop, considering the cost of hardware, software, and other expenses, amounts to about $ 1,000 per terabyte, or by one-fifth to a twentieth of the cost of other data management technologies (Giacalone, Scippacercola, 2016).

9. Benefits of Big Data in the analyzed fields

Benefits for e-Learning, and Learning Analytics. The Big Data are currently used by various companies for training and also in the university: with the help of Big Data, we can watch the learners and examine the traces of their individual paths. For example, we can identify the web pages on which learners will entertain more or which are more learning difficulties, those that often revisit, and determine the days and times they work out more, etc. Therefore, Big Data help us to understand the true role models for learners, much more than now occurs through the traditional education. These models lead to interesting information about what and how they learn. Thus, helping to make informed decisions about learning programs and to identify courses with design flaws. However, the real power of Big Data lies in their power to help predict or forecast scenarios to take preventive measures. For example, with the help of Big Data, it is also possible to make predictions as to what are the concepts that are difficult parts to students, the topics that generate confusion and difficulty in learning. Big Data today is presented as an effective platform that revolutionizes the traditional way in which e-learning was born. By Big Data it is possible to design more personalized learning plans and suitable for students. Monitoring becomes the main element used by educators. They are used the same results achieved by students to improve their training.

Benefits for Health Care. The Health Big Data are conceived as a real digital collection of all that the patient had, assumed or required in the medical field. The challenge, which is also the main difficulty of Big Data, is that the same should also merge any news on the patient's health that it communicates via social networks, like Facebook or Twitter, to his friends or acquaintances. In our life of patients each of us generates Big Data every time it receives a prescription, buy a drug, requires a health service, access to the ER, undergoing a diagnostic examination or laboratory uses Facebook, Twitter and other social network to communicate to friends their health. If it was a constant cross-analysis of all this information for each client: the doctors would have an overall picture of the health of the person, both in general and for a given period; policy makers, hospitals and clinics could include medical bills, preventing the most common diseases, and select healthcare services according to the real needs of the population in a given territory.For example, just an exchange of Tweet among people in constant contact to allow the data scientists to outline the possibility of infection and spread of a disease and then define the most appropriate prevention measures or to better manage health care demand. If the opportunities of Big Data will result in profiling of diagnostics and personalized therapies it will contribute to making the "health service" more effective and more sustainable also in financial terms (Roski et al., 2014).

Benefits for Banking Industry. Most of Big Data are useful for the customer management. The Bank's objective is to use this data to identify the customer profile more precisely (micro-segmentation) than with other methods. New services as tailored accounts to offer to customers are designed. The dialogue with consumers (sentiment analytics, multi-channel customer sentiment) is useful to identify the products that you could develop. The Bank can detect when a customer is about to leave the bank and it is possible to perform risk analysis better fraud detection more precise than before the advent of Big Data.

10. Conclusions

Nowadays we see the confluence of Big Data, Data Mining, Statistics, Mathematics, Computer Sciences, the Data Warehouse, the Artificial Intelligence, and neural networks in a new paradigm, which takes its name Data Science, and that promises to revolutionize the world, affecting all sectors, from health care, up to the academic world. In this perspective, *Data Science* will also modify the way of analyzing data.

The Data Science paradigm consists of extracting data of each type existing "in the world", applying appropriate formats, obtaining descriptive analysis of the phenomena, re-entering the results in the world circuit and so on, always perfecting more knowledge useful to the domain of the applications. From Data Science, it comes a new profession, the Data Scientist, who has the task of analyzing the data to provide useful information to make decisions to the customer.

The Data Scientist is the common professional figure for all analyzed sectors, drawing from the analysis of Big Data new and more effective strategies; he will have to learn to process information and be the responsible for statistical analysis, to determine what changes to make and to suggest choices to improve the processes.

Therefore, the role of Big Data is not only to be able to quickly handle large volumes of different data of various types, but is also given by the opportunity that these technologies offer us for new and remarkable applications, even in education, health care and banking.

With today's tools (mobile, tablets, smart phones, cloud-based technologies, etc.) the infrastructure is well-established. The Data Analytics allow us to get a much better picture of tracking than in the past with conventional methods used so far.

References

1. Alan, F. K., Sanil, A. P., and Sacks Arnold, K. E. **Signals: Applying Academic Analytics.** Educause Quarterly, Vol. 33, No. 1, 2010, pp. 1-10
2. Baker, R. S. and Inventado, P. S. **Educational data mining and learning analytics.** Learning Analytics, Springer, New York, 2014, pp. 61-75
3. Bohl, O., Scheuhase, J., Sengler, R. and Winand, U. **The sharable content object reference model (SCORM) a critical review.** Computers in education. Proceedings. International conference on IEE, 2002, pp. 950-995
4. Boinepelli, H. **Application of Big Data.** in Mohanty, H., Bhuyan, P. and Chenthati, D. (eds.) "Big Data: A Primer". Vol. 11. Springer, 2015
5. Chatti, M. A., Dyckhoff, A.L., Schroeder, U. and Thüs, H. **A reference model for learning ana-**

lytics. International Journal of Technology Enhanced Learning (IJTEL), Vol. 4, No. 5-6, 2012, pp. 318-331

6. Chawla, N. V., and Davis, D. A. **Bringing Big Data to personalized healthcare: a patient-centered framework.** Journal of general internal medicine, Vol. 28, No. 3, 2013, pp. 660-665
7. Clifton, B. **Advanced Web Metrics with Google Analytics,** 2nd ed., Wiley Publishing, Inc., Indianapolis, Indiana, 2010
8. Corposanto, C. and Lombi, L. **E-Methods and web society,** Università Cattolica del Sacro Cuore, Milano, 2014
9. Ferguson, R. **Learning Analytics: fattori trainanti, sviluppi e sfide.** TD tecnologie didattiche, Vol. 22, No. 3, 2014, pp. 138-147
10. Ferguson, R. **Learning Analytics: drivers, developments and challenges.** International Journal of Technology Enhanced Learning, Vol. 4, No. 5/6, 2012, pp. 304-317
11. Giacalone, M., and Scippacercola, S. **Il ruolo dei Big Data nelle strategie di apprendimento,** Atti Conferenza Didamatica, AICA ed., 2016, pp. 1-10
12. Groves, P., Kayyali, B., Knott, D., and Van Kuiken, S. **The 'Big Data'revolution in healthcare.** McKinsey Quarterly, Vol. 2, 2013
13. Gutierrez-Santos, S., Geraniou, S., Pearce-Lazard, S. D., and Poulovassilis, A. **Architectural Design of Teacher Assistance Tools in an Exploratory Learning Environment for Algebraic Generalisation.** IEEE Transactions of Learning Technologies, Vol. 5, No. 4, 2012, pp. 366-376
14. Hadoop, **http://Hadoop.apache.org/2014**.
15. Harrington, L. **Clinical intelligence.** Journal of Nursing Administration. Vol. 41, No. 12, 2011, pp. 507-509
16. Hipgrave, S. **Smarter fraud investigations with Big Data analytics,** Network Security, Vol. 12, 2013, pp.7-9
17. Kayyali, B., Knott, D. and Van Kuiken, S. **The big-data revolution in US health care: Accelerating value and innovation.** Mc Kinsey & Company, 2013, pp. 1-13
18. Kirkpatrick, D., L. **Techniques for evaluating training.** Training & Development Journal, Vol. 33, No. 6, 1979, pp. 78-92
19. Koza, J. R. **Genetic programming: on the programming of computers by means of natural selection,** vol 1. MIT Press: Cambridge, MA, 1992
20. Laney, D. **3D data management: Controlling data volume, velocity and variety.** Vol. 2, META Group Research Note, Vol. 6, 2001, p. 70
21. Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R. C., Roxburgh, C., and Byers, A. H. **Big Data: The next frontier for innovation,competition, and productivity.** McKinsey Global Institute, 2011
22. Manoochehri, M. **Data Just Right: Introduction to Large-Scale Data & Analytics.** Addison-Wesley Professional, 2013
23. Megahed, F. M., and Jones-Farmer, L. A. **Statistical Perspectives on "Big Data".** Frontiers in Statistical Quality Control, 11, 2015, pp. 29-47
24. Montgomery, D.C. **Introduction to Statistical quality control,** 7th edn.,Wiley, Hoboken, N.J., 2013
25. Murdoch, T. B., and Detsky, A. S. **The inevitable application of Big Data to health care.** Jama, Vol. 309, No. 13, 2013, pp. 1351-1352
26. Pappas, C. http://elearningindustry.com/Big-data-in-elearning-future-of-elearning-industry, 2014
27. Picciano, A. G. **The Evolution of Big Data and Learning Analytics in American Higher Education.** Journal of Asynchronous Learning Networks, Vol. 16, No. 3, 2012, pp. 9-20
28. Qiu, P. **Introduction to Statistical Process Control,** Boca Raton, FL: Chapman Hall/CRC, 2014
29. Raghupathi, W., and Raghupathi, V. **Big Data analytics in healthcare: promise and potential.** Health Information Science and Systems, Vol. 2, No. 1, 2014, p. 1

30. Reiss, C., Tumanov, A., Ganger, G. R., Katz, R. H., and Kozuch, M. A. **Heterogeneity and dynamicity of clouds at scale: Google trace analysis.** Proceedings of the Third ACM Symposium on Cloud Computing, ACM, 2012, p. 7
31. Rezzani, A. **Big Data: Architettura, tecnologie e metodo per l'utilizzo di grandi basi di dati.** Maggioli editore, 2013
32. Roski, J., Bo-Linn, G. W. and Andrews, T. A. **Creating value in health care through Big Data: opportunities and policy implications.** Health Affairs, Vol. 33, No. 7, 2014, pp. 1115-1122
33. Russom, P. **Big Data analytics.** TDWI Best Practices Report, Fourth Quarter, 2011, pp. 1-35
34. Sanchez, F. M., Gray, K., Bellazzi, R., and Lopez-Campos, G. **Exposome informatics: considerations for the design of future biomedical research information systems.** Journal of the American Medical Informatics Association, Vol. 21, No. 3, 2014, pp. 386-390
35. Scippacercola, S. **Metrics-based markov chains for web analytics.** Statistica & Applicazioni, Vol. 1, No. 12, 2012, pp. 55-66
36. Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S.B., Ferguson, R., Duval, E., Verbert, K. and Baker, R. S. J. D. **Open Learning Analytics: an integrated & modularized platform.** Doctoral dissertation, Open University Press, 2011
37. Slade, S., and Prinsloo, P. **Learning Analytics Ethical Issues and Dilemmas,** American Behavioral Scientist, Vol. 57, No. 10, 2013, pp. 1510-1529
38. Snijders, C., Matzat, U. and Reips, U. D. **Big Data: Big gaps of knowledge in the field of Internet.** International Journal of Internet Science, Vol. 7, 2012, pp. 1-5
39. Vaish G. **Getting Start with NoSQL,** Packt Publishing, 2013
40. Vasta, D. **Web Analytics,** Apogeo, Milano, 2009
41. Wright, A. H. **Genetic algorithms for real parameter optimization.** Foundations of genetic algorithms, Vol. 1, 1991, pp. 205-218
42. Zedlewski, J., Sobti, S., Garg, N., Zheng, F., Krishnamurthy, A. and Wang, R. Y. **Modeling Hard-Disk Power Consumption.** FAST, Vol. 3, 2003, pp. 217-230
43. Zong, W., and Wu, F. **The Challenge of Data Quality in the Big Data Age.** Journal of Xi'an Jiaotong University (Social Sciences), Vol. 33, No. 5, 2013, pp 38–43

[1] He is currently **Researcher in Statistics** and teaching staff member of the *Department of Economics and Statistics, University of Naples* "Federico II". Graduate in "Statistics and Economics Sciences", magna cum laude (Faculty of Economics – University of Palermo) he received his *PhD in Computational Statistics and Applications* from University of Naples "Federico II", Department of Mathematics and Statistics. During the PhD, he was *visiting scholar* of Prof. A.H. Money (Henley Management College). His research area encompasses the following topics: *Norm-p nonlinear regression - Multidimensional Data Analysis - Data Quality Control - Applications of Statistics in Medicine and in Justice*. Local component of the research groups funded by the University of Naples "Federico II" and co-financed by the relevant Ministry, he attended many Statistical Conferences organized by various national and international institutions, presenting numerous communications and papers.
Membership of "Società Italiana di Statistica", "International Association for Statistical Computing", "International Biometric Society" and "American Statistical Society", he gained over the years a considerable teaching experience as **Adjunct Professor** of "Statistics", "Probability", "Statistical Inference", "Medical Statistics", at various Italian Universities (Bologna, Naples "Federico II", Palermo, Catanzaro "Magna Graecia", Cosenza "University of Calabria", Messina, Catania). He is author of about eighty published works in Methodological and Applied Statistics.

[2] Associate **Professor of Statistics**, formerly **Professor of Information Processing Systems**, and he is a member of the teaching staff of the *Department of Economics, Management, and Institutions of the University of Naples* "Federico II". Graduate in Physics, he received the *Post-graduate in Theories and Techniques for the Use of Computers* from the Faculty of Engineering, University of Naples "Federico II". His research fields include *Multivariate Methods, Cluster Analysis, Business Intelligence and Decision Support Systems.*
He enjoys membership of the *American Statistical Society, Società Italiana di Statistica, CIRDIS, AICA* and of the *Working Group of the Italian Society of Statistic for Customer satisfaction and evaluation of services*.
He is Author of numerous publications in the Multivariate Statistical Analysis. He designed and developed models and algorithms for Decision Support Systems and for Ultra-Metrics.

ROBUST CONTROL CHARTS BASED ON MODIFIED TRIMMED STANDARD DEVIATION AND GINI'S MEAN DIFFERENCE

M.R. SINDHUMOL
Department of Statistics, University of Madras, India

M.R. SRINIVASAN
Department of Statistics, University of Madras, India

Michele GALLO
Department of Human and Social Sciences, University of Naples – L'Orientale, Italy

E-mail:

Abstract
Control Charts are process control techniques widely used to observe and control deviations and to enhance the quality of the product. Traditional control charts are based on the assumption that process data are independent in nature. Shewhart control charts are well known and are based on the basic assumption of normality. If process parameters are used to construct control limits based on preliminary samples, stability of the limits needs to be established as presence of outliers may affect the setting of control limits. In this paper an attempt has been made to first develop robust control charts based on trimmed mean and modified trimmed standard deviation. Secondly, an estimate of process standard deviation using Gini's Mean Difference (*G*) is also considered to modify the mean chart. Lastly, a comparative study is carried out to evaluate the performance of these two proposed robust charts with existing robust $\bar{x}$-MAD chart and two classical control charts namely $\bar{x}$-s chart and its modified $\bar{x}$-s_v chart, based on simulated data. Simulation study is also considered for performance evaluation of the proposed charts with other charts based on Average Run Length (ARL) and Operation Characteristic (OC) curves. In addition to the simulation, real data set is also used for setting up of robust control limits.

Keywords: *Control chart, Trimmed mean, Modified Trimmed Standard deviation, Gini's Mean Deviation, Outlier, Average Run Length, Operation Characteristic Curve*

1. Introduction

The Shewhart control chart, used for monitoring industrial processes is the most popular tool in Statistical Process Control (SPC), developed under the assumption of independence and normality. Performance of a control chart is dependent on the stability of the estimates used to construct control limits in phase I of the analysis. If selected estimators for constructing the limits are influenced by extreme values, setting up of limits may affect the control charting procedure. In situations, when the set limits are narrow, risk of a point falling beyond the limit increases thereby false indication of the process being out of control also gets increased. Similarly, if the limits are wider, the risk of points falling within the limits increases and hence falsely indicates the process to be in control.

When extreme values are present, mean and sample standard deviation are not considered to be good representatives of the data. A robust estimator is an estimator that is insensitive to changes in the underlying distribution and also resistant against the presence of outliers. Rockie (1989) suggested that in order to identify outliers, limits of a control chart be set based on robust measures while non-robust measures are plotted on it. There are many robust and non-robust measures of location and scale available in literature, which are used to develop control charts. The benefit of using control charts based on robust control statistics is that it does not have either a very high or a very low false alarm rate whenever the parameters to be controlled are close to the targets, although the data is no longer normal.

A location free, unbiased measure that can be used even with departure from normality is Gini's Mean Difference (*G*) introduced by Corrado Gini (1912). Yitzhaki (2003) introduced *G* as a superior measure of variability for non-normal distributions. Among the robust measures for dispersion, Median Absolute Deviation (*MAD*), introduced by Hampel (1974) is the widely used measure in various applications, as an alternative to sample standard deviation. Yitzhaki and Lambert (2013) showed that *MAD*, Least Absolute Deviation (*LAD*) and absolute deviation from a given quantile (*QUAD*) are actually either Gini's Mean Difference (G) of specific transformations applied to the distribution of the variables, or special cases of the between-group component of Gini's Mean Difference, called *BGMD*.

Riaz and Saghirr (2007) have developed a dispersion control chart based on G. Abu-Shawiesh (2008) introduced *MAD* to develop a robust dispersion control chart. Kayode (2012) used an estimate of process dispersion using *MAD* to improve mean chart and evaluated its performance through a comparative study of control charts.

The location charts use sample mean as an estimate for location parameter which are easily influenced by the extreme values and hence not suitable for heavy tailed distributions. The effect of extreme observations may be reduced with such observations being simply removed or given less weightage.

In this context, trimmed mean [Tukey, 1948] and its standard error are more appealing because of its computational simplicity. Apart from that, these measures are less affected by departures from normality than the usual mean and standard deviation. Wu and Zu (2009) showed that the trimmed mean is much more robust than its predecessor called Tukey trimmed mean. Relative to the mean, trimmed mean is highly efficient for large percentage of trimming at light-tailed symmetric distributions and much more efficient at heavy-tailed ones. Standard error of trimmed mean is not sufficient to estimate process dispersion because of trimming [Dixon and Yuen, 1974]. Huber (1981) obtained a jackknife estimator for its variance. Capéraà and Rivest (1995) derived an exact formula for variance of the trimmed mean as a function of order statistics, when trimming percentage is small. Sindhumol *et al.*(2015) modified trimmed standard deviation and observed it to be relatively more efficient compared to sample standard deviation. The γ-trimmed mean is defined as

$$\mu_t = \frac{1}{1-2\gamma}\int_{x_\gamma}^{x_{1-\gamma}} x\, f(x)dx. \qquad (1)$$

The trimmed mean is both location and scale equivariant. Its influence function is bounded but has jumps at x_γ and $x_{1-\gamma}$. It is qualitatively robust when $\gamma > 0$ and its breakdown point is γ. It shares the best breakdown point robustness of the sample median for any common trimming thresholds.

The principal purpose of this paper is to propose mean charts based on trimmed mean and its modified standard deviation. Moreover, simulation study is carried out to show the robustness at different levels of trimming. Further, a modified mean chart based on *G* is also developed as a robust control chart. Simulation study is carried out to compare the control limits of the proposed robust control charts with robust $\bar{\mathrm{x}}$-MAD chart and classical $\bar{\mathrm{x}}$-s chart and $\bar{\mathrm{x}} - s_v$ chart. Comparison of proposed charts performance using Operation Characteristic (OC) as well as Average Run Length (ARL) for different distributions is also carried out.

2. Classical Mean Charts

If x_{ij} $(i = 1,2,\dots,n$ and $j = 1,2,\dots,m)$ represents *m* random subgroups each of size *n* taken from a continuous and identical distribution with σ^2 unknown, an unbiased estimator is obtained using variance $s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$,where$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. If the process distribution is normal with parameters (μ, σ) control limits constructed for phase I analysis of mean chart are

$$\text{CL} = \bar{\bar{x}},\ \text{UCL} = \bar{\bar{x}} + 3\frac{\bar{s}}{(\sqrt{n})c_4} = \bar{\bar{x}} + A_3\bar{S},\ \text{LCL} = \bar{\bar{x}} - 3\frac{\bar{s}}{(\sqrt{n})c_4} = \bar{\bar{x}} - A_3\bar{S}. \qquad (2)$$

where *n* is the sample size and A_3 is a function of *n* and average of sample mean $\bar{x}$ and standard deviation *s* are taken over *m* subgroups.

Mahmoud *et al.* (2010) considered $\bar{S}_v = \left(\frac{1}{m}\sum_{i=1}^{m} S_i^2\right)^{1/2}$based on unbiased estimator of sample variance and showed that its efficacy in control charting application. Thus, using this estimator, a modification to classical location control limits can be set as

$$\text{UCL} = \bar{\bar{x}} + A\,\bar{S}_v;\ \text{CL} = \bar{\bar{x}};\text{LCL} = \bar{\bar{x}} - A\,\bar{S}_v. \qquad (3)$$

The above two standard classical control charts is chart is used to draw comparison and evaluate the performance of the proposed charts in the presence of outliers.

3. Robust chart based on Median Absolute Deviation

If $x_1, x_2, \dots, x_n$ is a set of observations, Hampel (1974) defined a robust estimate of dispersion as

$$MAD = 1.4826\ \text{median}\{|\mathrm{x_i} - \text{median}(\mathrm{x_i})|\}. \qquad (4)$$

Abu-Shawiesh (2008) used MAD to estimate and control process dispersion. Let *m* preliminary samples of size *n* are used to estimate, $\hat{\sigma} = b_n\overline{MAD}$, where average is taken over *m* subgroups. The constant b_n is the small sample correction factor given by Croux and Rousseeu (1992) as given below.

N	2	3	4	5	6	7	8	9
b_n	1.196	1.495	1.363	1.206	1.200	1.140	1.129	1.206

For n>9, b_n = n/(n-0.8). The general model for the control chart based on MAD is introduced by Abu-Shawiesh (2008). Kayode (2012) used MAD to modify limits of mean chart as

$$\mathrm{UCL} = \bar{\bar{x}} + \mathrm{A}_6\overline{MAD};\ \ CL = \ \bar{\bar{x}}\ \ ; \mathrm{LCL} = \bar{\bar{x}} - \mathrm{A}_6\overline{MAD} \qquad (5)$$

where $\bar{\bar{x}}$ is the average of the sample mean taken over *m* preliminary samples and $\mathrm{A}_6 = 3\ \frac{b_n}{\sqrt{n}}$. Adekeye and Azubuike (2012) have shown that this robust chart is good even for non-normal population.

4. Robust Control chart based on Gini's Mean Difference (G)

If $x_1, x_2, \ldots, x_n$ is a set of observations, an index of variability G is defined as

$$G = 2\sum_{j=1}^{n}\sum_{i=1}^{n}\left|x_i - x_j\right|/n(n-2) \quad (i \neq j) \qquad \text{and} = (\sqrt{\pi}/2)G\ . \qquad (6)$$

Riaz and Saghirr (2007) developed a dispersion control chart named *G*-chart based on G and its control limits are

$$\text{UCL} = \overline{K} + 3b_3\overline{K}; \quad \text{CL} = \overline{K}; \quad \text{LCL} = \overline{K} - 3b_3\overline{K}. \qquad (7)$$

where $\overline{K}$ is the average of K taken over *m* subgroups and b$_3$ is the standard deviation of G/σ.

A modified mean chart based on process dispersion estimated using G can be constructed with the control limits as

$$\mathrm{UCL} = \bar{\bar{x}} + A\ \overline{K}; \mathrm{CL} = \bar{\bar{x}}; \mathrm{LCL} = \bar{\bar{x}} - A\ \overline{K}. \qquad (8)$$

where $A = \ 3\ /\sqrt{n}$.

The proposed modified chart $\overline{x}$-K based on G is considered for performance evaluation with the classical $\bar{x} - sv$ chart as K is an unbiased estimator which can be used as an alternative to sample variance and s$_v$ is based on unbiased estimator.

5. Robust Control Charts based on trimmed mean

Langenberg and Iglewicz (1986) introduced trimmed mean $\bar{x}$ and R chart. Amin and Miller (1993) have developed robust $\bar{x}$ control chart usingVariable Sampling Interval (VSI) schemes and evaluated the behavior of VSI charts where trimmed mean, Winsorized mean, and the median are used. Figueiredo and Gomes (2004 and 2009) considered robust control charts based on the total median and on the total range statistics, for monitoring the process mean value and the process standard deviation, respectively. Schoonhoven.*et.al.* (2011) used a few robust location measures namely Median of means, 20% Trimmed mean of sample means, Hodges–Lehmann, Trimeans and 20% Trimmed mean of sample trimeans to develop mean charts. Schoonhoven and Does (2013) used Adaptively Trimmed Standard deviation denoted by ATS and 20%trimmed mean to improve mean chart.

Let x$_{(1)}$ ≤ x$_{(2)}$ …≤ x$_{(n)}$ denote an order statistics sample of size *n*, from a population having symmetric distribution. The *r*-times symmetrically trimmed sample is obtained by dropping both *r*-lowest and *r*-highest values. Here $r = [\alpha n]$ is the greatest integer and trimming is done for $\alpha\%(0 \leq \alpha \leq 0.5)$ of *n*. Trimmed mean is defined as

$$\bar{x}_T = \frac{\sum_{i=r+1}^{n-r} x_{(i)}}{n-2r}. \qquad (9)$$

Sample standard deviation of observations from trimmed mean is

$$s_T = \sqrt{\frac{\sum_{i=r+1}^{n-r}(x_{(i)} - \bar{x}_T)^2}{n-2r-1}}. \qquad (10)$$

Modified standard deviation is defined as $\widehat{\sigma}_{\mathrm{T}} = s_T^* = 1.4826\, s_T$ where the constant multiplier is used to cover the loss of information due to trimming. As percentage of trimming is increased, this constant gives a control on loss due to trimming. Sindhumol.*et al.*(2015) have introduced this modified standard deviation of trimmed mean, say s_T^* and have introduced it to control process dispersion. If process dispersion is estimated using this robust estimator, its average over m subgroups is, $\bar{s}_T^* = 1.4826\,\bar{\mathrm{s}}_\mathrm{T}$ and limits of dispersion chart are

CL $=c_4\bar{s}_T^*$; LCL = B$_3$ $\bar{s}_T^*$ and UCL = B$_4\bar{s}_T^*$. (11)

The constants B$_3$ and B$_4$ are the same constants as used in a classical chart. The limits of mean control chart can be modified using this modified standard deviation and depending upon the percentage of trimming, these limits have varying width as

CL = $\bar{\bar{x}}$; UCL = $\bar{\bar{x}} + A_3\bar{s}_T{}^*$; LCL = $\bar{\bar{x}} - A_3\bar{s}_T{}^*$. (12)

If location parameter is also estimated using trimmed mean, one can get control limits of robust mean chart for a particular level of trimming as

CL = $\bar{\bar{x}}_T$; UCL = $\bar{\bar{x}}_T + A_3\bar{s}_T{}^*$; LCL = $\bar{\bar{x}}_T - A_3\bar{s}_T{}^*$. (13)

The constants A$_3$ and A$_6$ are functions of sample sizes and the advantage of proposed control chart is the usage of the same constants A$_3$ as used in the classical control chart.

The proposed robust control charts based on two levels of trimming of 10% and 20%, namely $\bar{x} - s_T^*$ charts, $\bar{x}_T - s_T^*$, and the proposed robust control chart based on Gini, $\bar{x}$-*K* chart and the other robust chart $\bar{x}$-*MAD* discussed earlier along with the classical $\bar{x} - s$, $\bar{x} - s_v$ charts are considered for comparison.

6. Comparison of Robust charts for performance evaluation

The performance evaluation of the proposed robust charts is carried out as an empirical study based on Monte Carlo simulation conducted with 5000 runs using SAS software. Random samples generated from N(0,1) with sample sizes (n) of 10 and 20 with m=20 as number of samples for each cases are considered for simulation. Two levels of symmetric trimming at10% ($r = 1$) and 20% ($r = 2$) are considered for each subgroup. Clean samples are considered from N(0,1) to analyze type-1 error and contaminated samples are included to study the effect of detection of outliers or assignable causes of variation. The study considered out-of-control situation based on samples taken from N(1,1), N(2,1) and N(4,1). Samples from normal distribution with high location parameter, as is in the case of N(4,1), detection of shifts are easily studied in all the above charts. When small deviation in process location happened, as from samples N(1,1), even though false alarm is almost equal for all charts, charts based on 20% trimming detects contaminated sample more efficiently, compared to other charts. The performance of this chart is more clear and confirmed for the case of N(2,1). Simulation study also showed that the performance of mean chart is improved compared to the classical mean chart, when modified standard deviation of 20% trimming is used to estimate process dispersion.

A classical way of illustrating the effect of departure from normality is to consider contaminated normal distribution. Contamination of 40% is made for 5% of the data, which are the last four subgroups among 20. Hence sample mean is calculated so that 10% and 20% trimming will be meaningful for a contaminated subgroup. Thus, the contaminated models are 0.60 N(0,1) with 0.40 N(1,1) and 0.40 N(2,1).

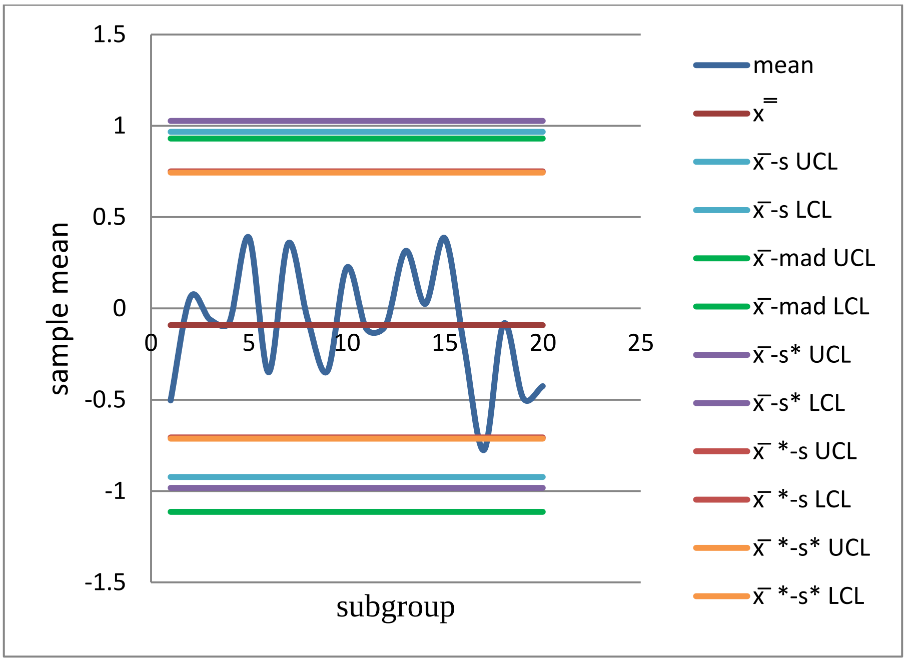


Figure 1: Control limits of modified trimmed mean charts, $\bar{x}$-s_T^* and $\bar{x}_T$-s_T^* and other charts based on MAD and s with n=10, m=20 and 10% contamination using N(2,1)

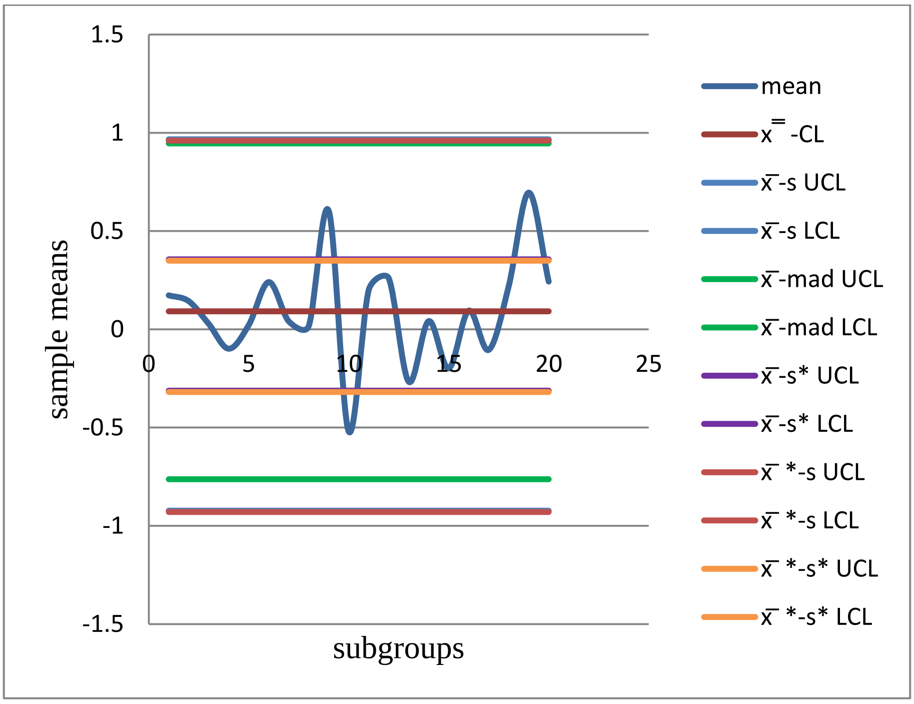


Figure 2: Control limits of modified trimmed mean charts, $\bar{x}$-s_T^* and $\bar{x}_T$-s_T^* and other Charts based on MAD and s with n=10, m=20 and 20% Contamination using N(1,1).

When charts $\bar{x}$-s, $\bar{x}$-s_v, $\bar{x}$-MAD, $\bar{x}$-s_T^*(with 10% and 20% trimming) and $\bar{x}_T$-s_T^*(with 10% and 20% trimming) are compared in presence of contamination, more contaminated samples are detected by $\bar{x}_T$-s_T^* chart than other charts having almost same type I error. Also, for 10% level of trimming, charts $\bar{x}_T$-s and $\bar{x}_T$-s_T^* perform equally well and have the ability to detect small variations too. Though the chart $\bar{x}$-$s_T^*$is bit wider compared to $\bar{x}$-s chart due to the effect of multiplier, chart $\bar{x}_T$-s$_T^*$ has smaller width compared to $\bar{x}$-MAD. Figure-1 shows performance of modified trimmed mean chart compared to other charts including chart based on MAD, in terms of detection of samples when contaminated with N(2,1) ($\bar{x}_T = \bar{x}^*$and s$_T^*$ = s*in all the figures).

At 20% level of trimming, width of the charts reduces respectively for charts $\bar{x}$-s, $\bar{x}$-MAD and $\bar{x}$-s$_T^*$and performance of charts $\bar{x}$-s$_T^*$ and $\bar{x}_T$-s$_T^*$ are same. Increase in trimming makes loss of data as well as an increase in type I error. Still these charts have the ability to detect smaller variation and can even act as warning limits in case of larger trimming levels. Selection of dispersion measure influences performance of mean chart and the simulation study shows that s$_T^*$is a better choice. Figure-2 shows performance of the modified trimmed mean chart compared to other charts including chart based on MAD, in terms of detection of samples when contaminated with N(1,1) respectively. The charts $\bar{x}$-s$_T^*$ and $\bar{x}_T$-s$_T^*$ have same and smaller width which helped in the early detection of variations.

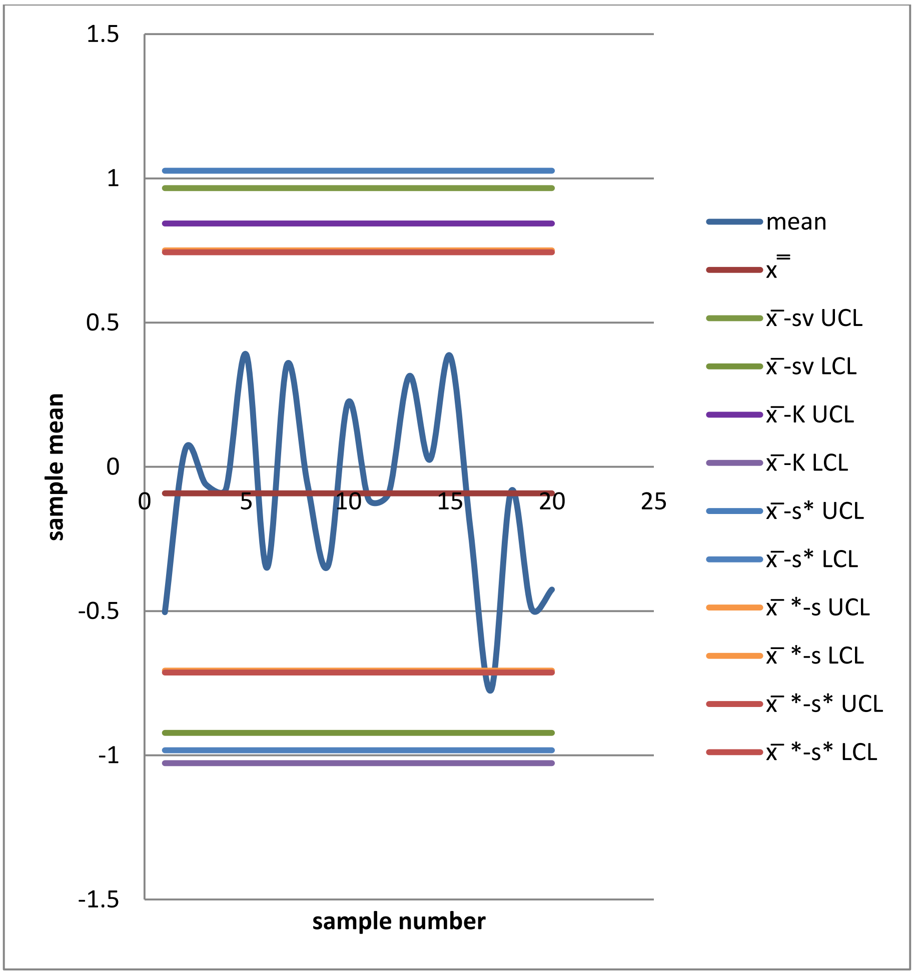


Figure 3: Control limits of modified trimmed mean charts, $\bar{x}$-s$_T^*$ and $\bar{x}_T$-s$_T^*$ and $\bar{x}$-K Charts with other charts based on MAD and s$_v$ with n=10, m=20 and 10% contamination using N(2,1).

When charts $\bar{x}$-s$_v$, $\bar{x}$-K, $\bar{x}$-s$_T^*$ (with 10% and 20% trimming) and $\bar{x}_T$-s$_T^*$ (with 10% and 20% trimming) are compared in presence of contamination, more contaminated samples are detected by $\bar{x}_T$-s$_T^*$ chart than other charts having almost same type I error. At 10% level of trimming, chart $\bar{x}$-s$_T^*$showed almost same width as $\bar{x}$-s$_v$ due to the effect of multiplier s$_T^*$. The charts with smaller width namely, $\bar{x}_T$-s and $\bar{x}_T$-s_T^* perform equally well and have ability to detect small variations too. Mean chart with dispersion measure K is a good alternative to that of s_T^* as width of $\bar{x}$-K chart is smaller in width than that of $\bar{x}$-$s_T^*$and $\bar{x}$-s_v, for small percentage of trimming. Figure -3 showed performance of modified trimmed mean chart compared to other charts in terms of detection of samples when contaminated with N(2,1).

At 20% level of trimming, chart $\bar{x}$-s_T showed almost same performance and width as $\bar{x}_T$-s_T^* and have the smallest width among other charts. All other charts, including charts based on K, have almost the same width. Mean chart with dispersion measure K is a good alternative to mean chart with s_v, though both are based on unbiased estimators. Figure -4 shows performance of trimmed mean chart compared to other charts including $\bar{x} - K$ in terms of detection of samples when contaminated with N(1,1).

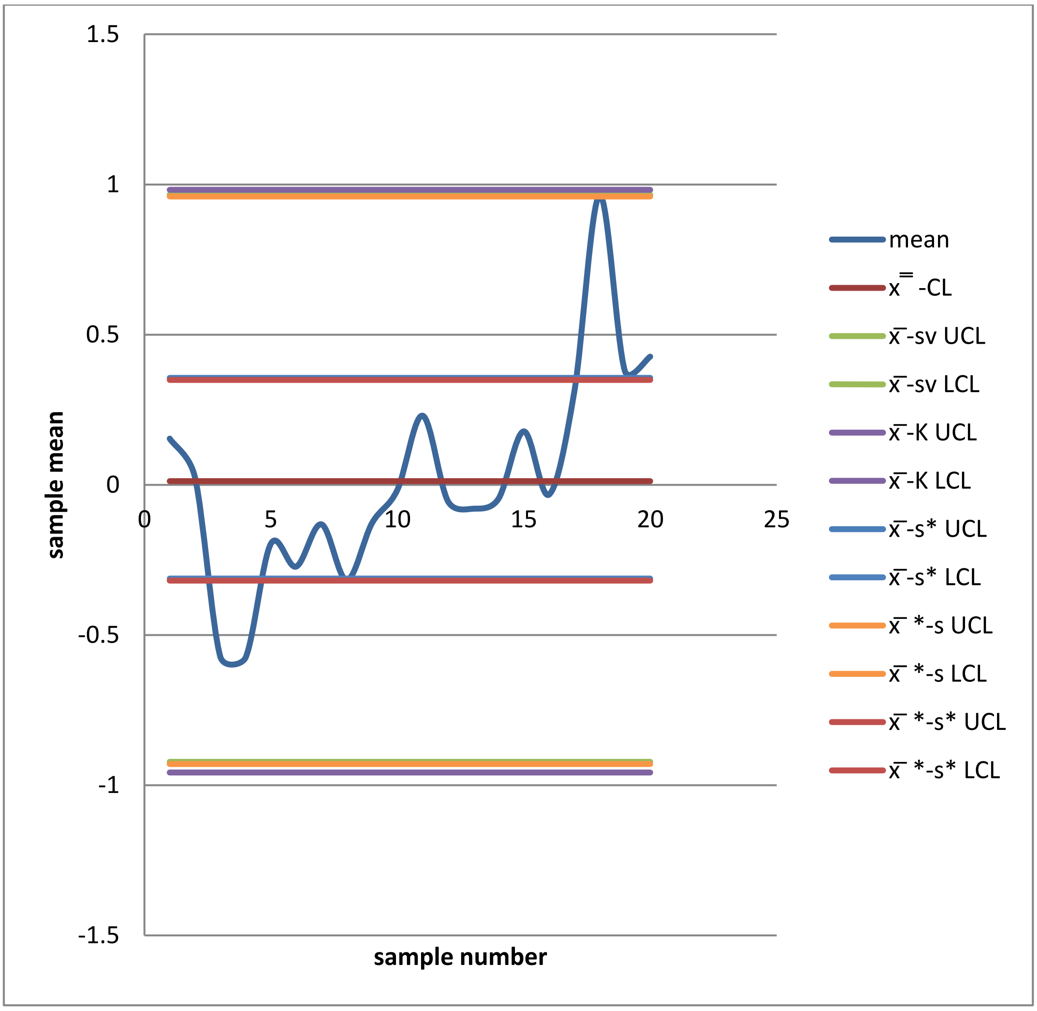


Figure 4: Control limits of modified trimmed mean charts, $\bar{x}$-s_T^* and $\bar{x}_T$-s_T^* and $\bar{x}$-K Charts with other charts based on MAD and s_v with n=10, m=20 and 20% contamination using N(1,1).

6.1. Performance using Operating Characteristic (OC) curves

The ability of the chart to detect shift in quality due to assignable causes of variation is accessed by Operating Characteristic curve. The OC curve of the $\bar{x}$-chart for phase II analysis is studied here. The OC function of $\bar{x}$–chart, is the probability of not detecting the shift in the process mean μ on the first subsequent sample of size n taken after the shift has happened. If the mean in the in-control state is μ_0 and the shift is $\mu_1 = \mu_0 + k\sigma$, β-risk is defined under standard normal distribution function φ as

$$\beta = P(LCL \leq \bar{x} \leq UCL \,/\, \mu = \mu_1) = \varphi\,(3 - k\sqrt{n}\,) - \varphi\,(-3 - k\sqrt{n}\,). \qquad (14)$$

The OC curve for the chart is drawn by plotting the β -risk against the magnitude of the shift in quality which is expressed in standard deviation units for various sample sizes *n*. Process s.d is considered to be known or to be estimated before considering β–risk. Steeper the curve better the probability of detection of shift, which is variation in the process quality. Figure 5 shows that charts $\bar{x}_T$-s and $\bar{x}_T$-s_T^* have the smallest β-risk compared to all other charts. Mean charts modified with robust measures MAD and G have almost same β-risk and OC curves are overlapping.

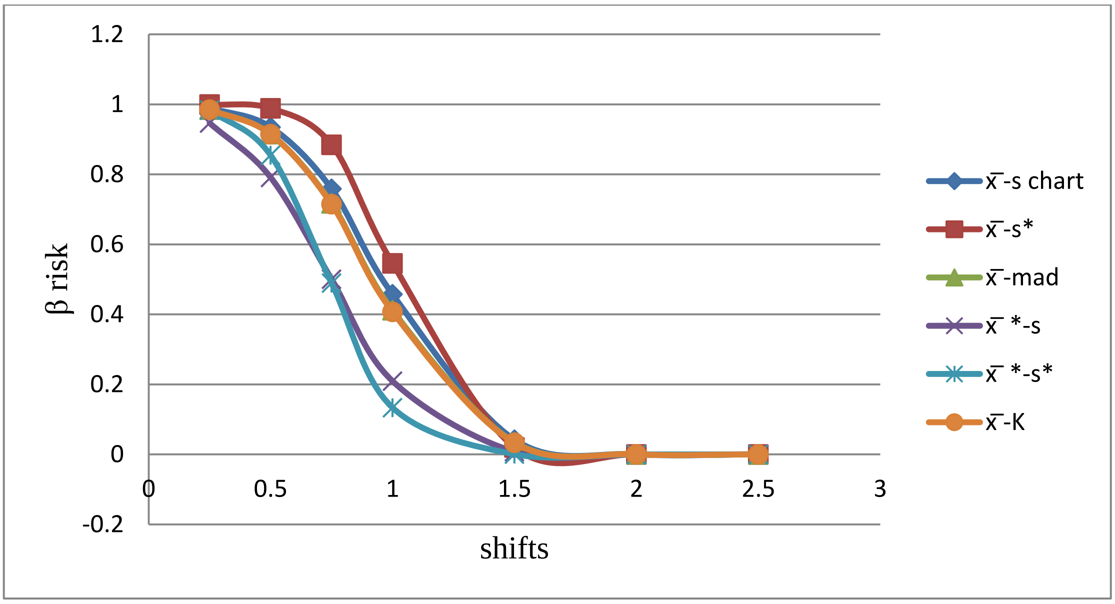


Figure 5: Comparison of OC curves for modified trimmed mean charts and $\bar{x}$-K chart with other charts based on s and MAD.

6.2. Performance Using Average Run Length

Average Run Length is defined as reciprocal of the probability that any point exceeds the control limits. It shows how often false alarms occur or how often the chart detects a shift in quality. If the process is in-control, Average Run Length, ARL_0, should be large and if the process is out-of-control ARL should be very small. The number of subgroups taken before an out-of-control $\bar{x}$ happened is recorded as a run length observation, RLi. The process is repeated 10000 times and the results of this simulation study are given in Table 1. The ARL_0 was calculated as $ARL_0 = \frac{\sum_{i=1}^{10000} AR_i}{10000}$. The same procedure is used to compare the out-of-control ARL_1, that is, in presence of assignable causes also.

Table 1: ARL for sample size n=10 and m=20

Distributions	$\bar{x} - s$ **chart**	$\bar{x}$ −**S**var **chart**	$\bar{x}$ −**MAD chart**	$\bar{x} - K$ **chart**	$\bar{x}$ −**S**T **chart**		$\bar{x}_T$ −**S**T **chart**	
					r=1	**r=2**	**r=1**	**r=2**
N(0,1)	323	322	319	230	506	64	527	63
N(1,1)	544	544	581	620	470	0.283	470	0.275
N(2,1)	0.0008	0.0008	0.006	0.004	0.0008	0	0.0008	0
N(0,1)+40%	317	317	331	441	256	0.74	261	0.72

N(1,1)								
N(0,1)+40% N(2,1)	66	66	68	83	65	6.43	65	6.14

Table 1 shows that among the mean charts, dispersion estimated using s_T^* with 10% trimming ARL$_0$ calculated on N(0,1) is larger than that of other charts. Hence small percentage of trimming can be advisable even for a controlled process. In presence of assignable causes, ARL$_1$ is better for 20% trimming. The ARL comparison supports mean chart modified with robust MAD than that of G. The study shows the need of using robust measure of dispersion in place of sample standard deviation, especially s_T^* while using mean chart to control a process.

The ARL study supports MAD as a better choice as compared to G to modify mean charts. The study has established use of trimmed mean and modified trimmed standard deviation to estimate process mean and standard deviation and mean charts modified on these estimate as better choices in quality control applications.

7. Charts performance on real data

The performance of charts based on trimmed means at two levels has shown to be a better performer compared to other robust charts and classical charts based on simulation studies, OC curves and ARL. The performance of the charts is validated on real data shown in Figure 6.

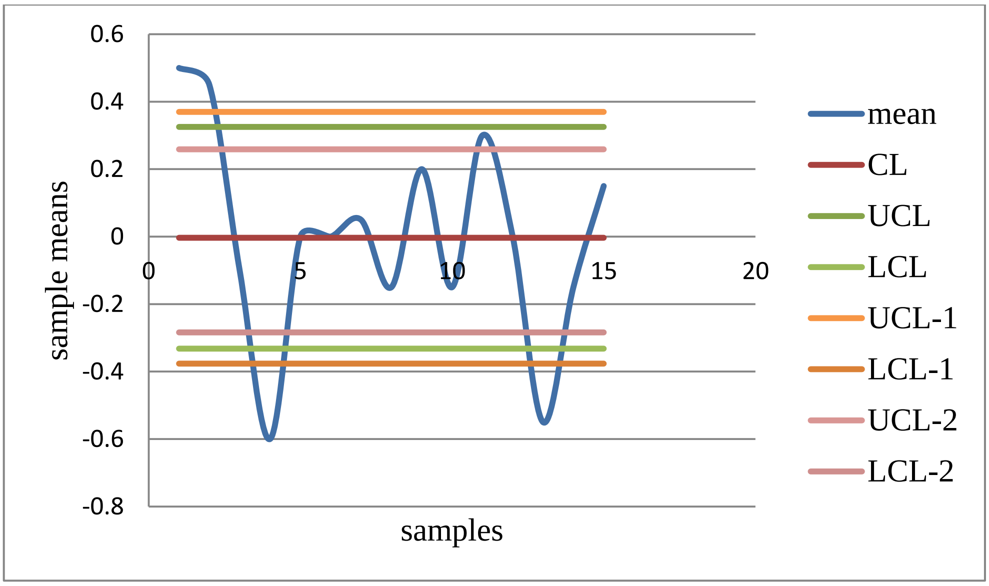


Figure 6: Comparison of modified trimmed mean charts for trimming 10% and 20% respectively with classical mean chart based on data from D.C.Montgomery, Table 6E.5

The fill volume of soft drink beverage bottles is considered as a quality characteristic. The volume is measured (approximately) by placing a gauge over the crown and comparing the height of the liquid in the neck of the bottle against a coded scale. On this scale a reading of zero corresponds to the correct fill height. Fill heights are shown by Montgomery (2009) in Table 6E.5 and fifteen samples of size n=10 have been analyzed using mean charts with dispersion measures s (in figure LCL, UCL), modified trimmed standard deviation s_T^* for 10%

(in figure LCL-1, UCL-1) and for 20% (in figure LCL-2, UCL-2). The s_T^*-chart for 20% trimming detected 10th sample also as out of control point.

8. Conclusion

Trimmed mean and standard deviation based on it are robust measures of location and scale. Trimmed mean is already exposed to control charting process. However, actual variance of trimmed mean which is a function of order statistics or variance of trimmed mean modification based on Winsorization, are not helping to represent process variance. The modified standard deviation presented by Sindhumol.et.al. (2015) is a better alternative in this regard.

Selection of dispersion measure as well as location measure to estimate process parameters influence performance of mean chart. In this study two robust location control charts, one based on measures on trimmed data and the other based on G are proposed. Four types of trimmed mean charts including trimmed mean (10% and 20% of trimming), modified trimmed standard deviation for two levels of trimming (10% and 20% of trimming), two robust mean charts one modified with G and the other with MAD, two classical mean charts based on unbiased sample variance and biased sample standard deviation, are considered and a simulation study is conducted for comparing performance of charts in terms of false and correct detection of outliers. In presence of assignable causes of variation, charts based on trimmed mean and modified trimmed standard deviation, the results are outstanding even for small shifts. The chart can be even used as warning limits for early detection of assignable causes, for large trimming percentage. The chart $\bar{x}_T$-s_T^* has smaller width compared to $\bar{x}$-MAD and $\bar{x}$-K charts, especially in large trimming. Mean chart with dispersion measure K is a good alternative to that of s_v (Mahmoud et al. , 2010) though both are based on unbiased measures of dispersion, as width of $\bar{x}$-K chart is smaller. The study shows that to modify to mean chart, MAD is slightly a better choice than G. It is noted that, both $\bar{x}$-s_T^* and $\bar{x}_T$-s_T^* (at10% trimming) perform better than all other charts when there is no contamination. The comparative study of limits in terms of false detection shows that all charts are alike except for 20% trimming.

The OC curves show that charts $\bar{x}_T$-s and $\bar{x}_T$-s_T^* have the smallest β-risk compared to all other charts and hence have the largest power to detect assignable causes of variations. Mean charts modified with robust measures G and MAD have almost same β-risk.

The ARL study shows that a small percentage of trimming is advisable even for an in-control process. Robust chart based on trimmed mean and modified trimmed standard deviation chart excel in performance for small percentage of trimming and can even be used as warning limits for large percentage of trimming, in the presence of assignable causes. The ARL comparison supports mean chart modified with robust MAD than that of G while OC curve study gives all most equal β-risk.

Robust location chart based on trimmed mean and modified trimmed standard deviation can be used as an effective tool to control and improve process performance. This robust limits help to detect assignable causes or outliers in the phase II analysis if statistical control of process location parameter is tested using sample means collected. Both ARL and OC study show that, $\bar{x}$-$s_T^*$and $\bar{x}_T$-s_T^* (at10% trimming) perform better than all other charts under contamination or otherwise. The study also that in order to frame warning limits, one can increase the trimming percentage. Hence study supports usage of data with small percentage of trimming to frame control limits in phase I analysis to get a better control in phase II analysis of a process.

Charts are tested on real data also, with two levels of trimming. The robust chart for controlling process location parameter using trimmed mean and modified trimmed standard deviation is a better option than classical mean chart in real data too. This modified trimmed mean chart with 20% trimming identifies more outlier points.

Study shows that the performance of classical mean chart is increased if process dispersion is estimated using s_v and the robust measure G is a good alternative to $s_{v.}$ Among the G and MAD, for constructing a robust control chart, MAD seems to be a better estimate but trimmed mean and modified trimmed standard deviation are better pairs of estimates in quality control applications.

References

1. Abu-Shawiesh, M. O. A. **A simple Robust Control Chart based on MAD**, Journal of Mathematics and Statistics, Vol.4, No.2, 2008, pp.102-107
2. Adekeye, K .S. and Azubuike, P. I. **Derivation of the Limits for Control Chart Using the Median Absolute Deviation for Monitoring Non-Normal Process**, Journal of Mathematics and Statistics, Vol. 8, No.1, 2012, pp. 37-41
3. Amin, A. W. and Miller, R. W. **A Robust study of X-bar chart with variable sampling interval**, Journal of Quality Technology, Vol. 25, No.1, 1993, pp.36-44
4. Caperaa, P. and Rivest, L. P. **On the variance of the trimmed mean**, Statistics & Probability Letters, Vol. 22, 1995, pp. 79-85
5. Ceriani, L. and Verme, P. **The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by CorradoGini,** The Journal of Economic Inequality*,* Vol. 10, No. 3, 2012, pp. 421-443
6. Croux, C. and Rousseeuw, P. J. **Time efficient algorithms for two highly robust estimators of scale,** Computational Statistics, Vol.1, 1992, pp. 411-428
7. Dixon, W. J. and Yuen, K. K. **Trimming and Winsorization: A review,** Statistische Hefte, Vol. 15, No. 2,1974, pp. 157-170
8. Figueiredo, F. and Gomes, M. I. **The total median is Statistical Quality Control,** Applied Stochastic Models in Business and Industry, Vol. 20, 2004, pp. 339–353
9. Figueiredo, F. and Gomes, M. I. **Monitoring Industrial Processes with Robust Control Charts,** REVSTAT – Statistical Journal, Vol. 7, No. 2, 2009, pp. 151–170
10. Hampel, F. R. **The influence curve and its role in robust estimation,** Journal of the American Statistical Association, Vol. 69, 1974, pp. 383- 393
11. Huber, P.J. **Robust Statistics**, John Wiley, New York, 1981
12. Kayode, A. S. **Modified Simple Robust Control Chart based on Median Absolute Deviation,** International Journal of Statistics and Probability, Vol. 2, No. 4, 2012, pp. 91-95
13. Langenberg, P.and Iglewicz, B. **Trimmed Mean X bar and R chart**, Journal of Quality Technology, Vol. 18, 1986, pp. 152-161
14. Mahmoud, A. M., Henderson, G. R., Epprecht, E. K. and Woodall, W. H. **Estimating the standard deviation in Quality control applications,** Journal of Quality Technology, Vol. 42, No. 4, 2010, pp. 348-357
15. Riaz, M. and Saghirr, A. **Monitoring Process Variability Using Gini's Mean Difference,** Quality Technology and Quantitative Management, Vol. 4, No. 4, 2007, pp. 439-454
16. Rockie, D. M. **Robust Control Charts**, Technometrics, Vol.31, 1989, pp. 173-184
17. Schoonhoven, M., Nazir, H. Z., Riaz, M. and Does, R. J. M. M. **Robust location estimators for x bar control charts,** Journal of Quality Technology, Vol. 43, No. 4, 2011, pp. 363-379

18. Schoonhoven, M. and Does, R. J. M. M. **A robust X bar control chart,** Quality and Reliability Engineering International, Vol. 29, 2013, pp. 951–970
19. Sindhumol, M. R., Srinivasan, M. R. and Gallo, M. **A robust dispersion control chart based on modified trimmed standard deviation,** Electronic Journal of Applied Statistics, Vol. 9, No. 1, 2016, pp. 111-121
20. Tukey, J. W. **Some elementary problems of importance to small sample practice,** Human Biology, No. 20, 1948, pp. 205–214
21. Yitzhaki. S. **Gini's Mean difference: a superior measure of variability for non-normal distributions,** METRON - International Journal of Statistics, vol. LXI, No. 2, 2003, pp. 285-316
22. Yitzhaki, S. and Lambert, P. J. **The Relationship between Gini's Mean Difference and the Absolute Deviation from a Quantile**, METRON, Vol. 71, 2013, pp. 97–104
23. Yohai, V. J. and Zamar, R. H. **High Breakdown Point Estimates Regression by Means of the Minimization of an Efficient Scale,** Journal of the American Statistical Association, Vol. 83, 1988, pp. 406-413
24. Wu, M. and Zuo, Y. **Trimmed and Winsorized means based on a scaled deviation**, Journal of Statistical Planning and Inference, Vol. 139, 2009, pp. 350- 365

THE PERFORMANCE OF THE SRMR, RMSEA, CFI, AND TLI: AN EXAMINATION OF SAMPLE SIZE, PATH SIZE, AND DEGREES OF FREEDOM

Gita TAASOOBSHIRAZI[1]
Associate Professor, Department of Quantitative and Mixed Methods Research Methodologies, University of Cincinnati, USA

E-mail: gita.tshirazi@gmail.com

Shanshan WANG[2]
PhD Student, Department of Quantitative and Mixed Methods Research Methodologies, University of Cincinnati, USA

Abstract

The SRMR, RMSEA, CFI, and TLI are commonly used fit indexes reported when describing the fit of structural equation models (SEM) used in math and science education. A large number of the models tested in math and science education tend to be path models that study the interaction between various motivational, affective, contextual, and cognitive variables or latent growth curve models that examine change in students over time. The majority of these models tend to have small degrees of freedom and small sample sizes. Given the common use of these fit indexes, it is important to understand their performance when reported for relatively simple models.

Keywords: *structural equation modeling; growth model; sample size; degrees of freedom; simulation; science and math education*

The standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), comparative fit index (CFI), and the Tucker Lewis Index (TLI) are commonly used fit indexes reported when describing the fit of structural equation models (Kline, 2010; Worthington & Whittaker, 2006), with the RMSEA, SRMR, and CFI being among the most widely reported in the SEM literature (Kline, 2010).

The SRMR is a measure of the mean absolute correlation residual, with smaller values suggesting good model fit (Kline, 2010). The RMSEA provides information about 'badness of fit', with lower RMSEA values indicating good model fit (Kline, 2010). The CFI and TLI are both incremental fit indexes that assess the improvement in the fit of a model over that of a baseline model with no relationship among the model variables; larger values indicate better model fit (Kline, 2010).

Several studies have examined the performance of these and other fit indexes under various conditions (e.g., impact of sample size on fit index values) using simulations (Chen et al., 2008; Hu & Bentler, 1999). However, this work has examined the performance of the various fit indexes under different conditions with models that have moderate to large degrees of freedom. For example, (Hu & Bentler, 1999) examined the performance of the SRMR, TLI, CFI, and RMSEA with models with degrees of freedom larger than 80. There is a dearth of research examining the performance of SEM fit indexes using models with smaller degrees of freedom typical of path models or latent growth curve models tested in math and science education (Kenny, Kaniskan & McCoach, 2014; Kenny & McCoach, 2003).

Studies that have examined the impact of degrees of freedom (Breivick & Olsson, 2001; Kenny & McCoach, 2003) on model fit have tended to focus on the RMSEA and have found that the RMSEA showed better fit (smaller RMSEA values) for models with larger degrees of freedom. Kenny, Kaniskan, and McCoach (2014), who studied the performance of the RMSEA with models with small degrees of freedom, found that models with a combination of a smaller degrees of freedom and smaller sample sizes had RMSEA values that often falsely indicated a poor model fit. The authors found the RMSEA to be elevated with small sample sizes (N ≤ 100). As the model degrees of freedom decreased, model rejection rates increased for the RMSEA, even with sample sizes as large as 1000. The RMSEA decreases if there are more degrees of freedom and/or a larger sample size, keeping everything else constant (Kline, 2010). This suggests that more parsimonious models have smaller RMSEA values. However, with the exception of the Kenny, Kaniskan, and McCoach (2014) study that looked at 1, 2, 3, 5, 10, 20, and 50 degrees of freedom with sample sizes that ranged from 50 to 1,000, other studies examining the RMSEA have used degrees of freedom much larger (e.g., 24 to 528 degrees of freedom) than those found in models frequently tested in math and science education.

The SRMR, RMSEA, CFI, and TLI are commonly used fit indexes reported when describing the fit of models used in math and science education (e.g., Bailey, Taasoobshirazi, & Carr, 2014; Byars-Winston & Fouad, 2008; Mettern & Schau, 2002; Stevens, Olivarez, Lan, & Tallent-Runnels, 2004; Glynn, Taasoobshirazi & Brickman, 2007). Many of the models tested in math and science education are path models that examine the interaction between various motivational, affective, cognitive, and contextual variables (e.g., Ha, Haury, & Nehm, 2012; Kingir, Tas, Gok, Vural, 2013; Kirbulut, 2014; Stevens et al., 2004). These models tend to have fewer than 10 variables and small degrees of freedom (less than 10) (e.g., Adedokun, Bessenbacher, Parker, & Kirkham, 2013; Akyol, Tekkaya, Sungur, & Traynor, 2012; Bailey et al., 2014). In addition, a common structural equation model with a small degrees of freedom tested in math and science education is the latent growth model (e.g., Gottfried, Marcoulides, Gottfried, & Oliver, 2009). Given the common use of these fit indexes, it is important to understand their performance when used with models with small degrees of freedom.

Also of interest was the impact of sample size on the performance of the fit indexes with models with small degrees of freedom. For example, the research in math and science education testing path models often have sample sizes that tend to be less than N = 250, with many studies having sample sizes less than N = 100 (e.g., Bailey et al., 2014; Ha, Haury, & Nehm, 2012). This is much smaller than the sample sizes typically found in simulation studies on fit indexes in SEM (e.g., Hu & Bentler, 1999). Several studies have shown the importance of sample size on the performance of fit indexes such as the RMSEA (Chen et al., 2008; Kenny et al., 2014). Specifically, Type II error rates for the fit indexes increase as sample size decreases. A goal of the present study was to determine the interaction between sample size and degrees of freedom on the performance of the SRMR, RMSEA, CFI, and TLI for models that have small degrees of freedom and small sample sizes, typical of what is found in the path models and latent growth curve models tested in math and science education.

We wanted to know, for our four fit indexes and when working with models with small degrees of freedom: What is the performance of these fit indexes and their rejection rates across various sample size and degrees of freedom combinations? Specifically, do models with smaller degrees of freedom (more paths in the model) require a larger sample size, similar to the results of the Kenny, Kaniskan, and McCoach (2014) findings for the RMSEA or is a smaller sample size sufficient for models that have smaller degrees of freedom in line with research illustrating that adding paths to a model tends to improve fit (Kline, 2010)? What is a sufficient sample size for a small degrees of freedom model needed to avoid making a type II error?

Method

We conducted a Monte Carlo simulation to test correctly specified growth models with varying sample sizes and degrees of freedom. To do so, we followed the same simulation techniques as Kenny et al. (2014). Specifically, "intercept loadings were all fixed to one and slope loadings were fixed to zero for wave 1 and increased in one-unit increments thereafter. The population mean of the intercept factor was 0.5 and the variance was set at 1.0: The population mean of the slope factor was 1.0 and its variance was 0.2. The covariance between slope and intercept was 0.1, and all error variances were set at 0.5. The models were as follows based and are designated by their degrees of freedom: df = 1: 3-wave growth model, df = 2: 3-wave growth model with equal error variances and the loading for the slope factor at wave 3 free, df = 3: 3-wave growth model, with equal error variances, df = 5: 4-wave growth model, df = 10: 5 wave growth model (figure 1), df = 20: 7 wave growth model, with loadings on the slope factor for the last three times free, and df = 50: 10 wave growth model" (Kenny et al., 2014, p. 10). In addition, we used the same sample sizes as Kenny et al., (2014) including: 50, 100, 200, 400, 600, and 1,000.

The simulation feature of Mplus Version 7.3 was used to generate the data, and the R package MplusAutomation (Hallquist & Wiley, 2014) was used to estimate the parameters for each simulation condition. We replicated each condition 1,000 times. The raw data were generated from a multivariate normal distribution. In order to address our research questions, we calculated the average value of the model fit indices for each simulation condition and their accompanying rejection rates defined as the number of times that the model fit indices exceeded recommended cutoff values specified in the research. Specifically, a CLI cutoff value of .95, a TLI cutoff value of .95, a RMSEA cutoff value of .05, and a SRMR cutoff value of .08 were used (Hooper, Coughlan, & Mullen, 2008).

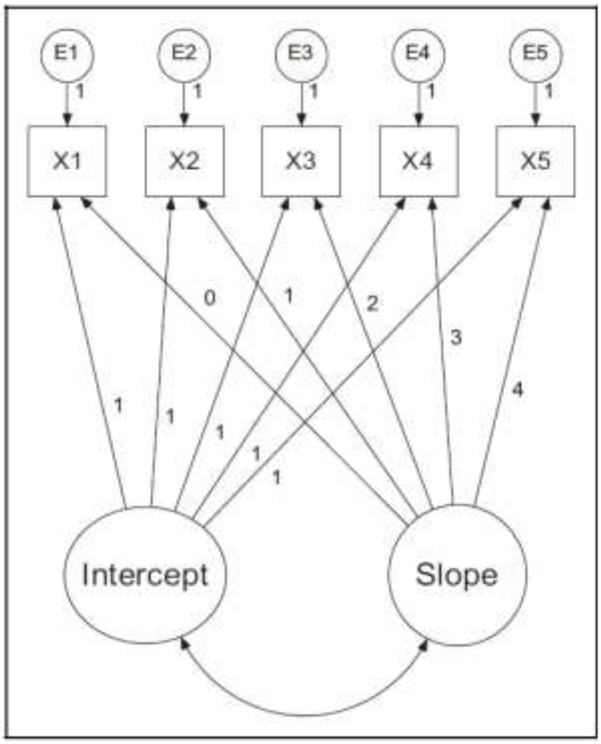


Figure 1. Simulation model for 10 degrees of freedom.

Results

CFI. Results of the simulation for the various degrees of freedom and sample sizes are presented in Table 1. The table reports CFI values and accompanying rejection rates. Figure 2 shows the CFI values for the selected sample sizes and degrees of freedom. Results indicated that adding paths to the model (decreasing the degrees of freedom) did not do much in terms of altering CFI values. Increasing sample size resulted in larger CFI values. Based on rejection rates, we almost never found the CFI to fall below the cutoff for sample sizes 100 or larger. For N = 50, the largest rejection rate was about 10%, suggesting that such small sample sizes may be problematic when using the CFI (unless degrees of freedom are very large).

Table 1. CFI values for selected degrees of freedom and sample size.
Rejection rates are in parentheses

df	1	2	3	5	10	20	50
N = 50	0.992 (3.90%)	0.990 (6.00%)	0.986 (9.80%)	0.990 (5.10%)	0.991 (2.30%)	0.989 (3.20%)	0.992 (0.10%)
N = 100	0.996 (0.30%)	0.995 (1.60%)	0.993 (2.20%)	0.995 (0.20%)	0.996 (0.10%)	0.995 (0.10%)	0.997 (0.00%)
N = 200	0.998 (0.00%)	0.997 (0.20%)	0.996 (0.20%)	0.997 (0.00%)	0.998 (0.00%)	0.998 (0.00%)	0.999 (0.00%)
N = 400	0.999 (0.00%)	0.999 (0.00%)	0.998 (0.00%)	0.999 (0.00%)	0.999 (0.00%)	0.999 (0.00%)	0.999 (0.00%)
N = 600	0.999 (0.00%)	0.999 (0.00%)	0.999 (0.00%)	0.999 (0.00%)	0.999 (0.00%)	0.999 (0.00%)	1.000 (0.00%)
N = 1000	1.000 (0.00%)	1.000 (0.00%)	0.999 (0.00%)	0.999 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)

TLI. Results of the simulation for the various degrees of freedom and sample sizes are presented in Table 2.

Table 2. TLI values for selected degrees of freedom and sample size.
Rejection rates are in parentheses

df	1	2	3	5	10	20	50
N = 50	0.997 (16.10%)	1.001 (12.20%)	0.998 (9.80%)	0.998 (7.30%)	0.998 (2.30%)	0.995 (3.70%)	0.995 (0.10%)
N = 100	0.999 (8.30%)	0.999 (4.00%)	0.999 (2.20%)	0.999 (0.90%)	0.999 (0.10%)	0.999 (0.10%)	0.999 (0.00%)
N = 200	0.999 (2.10%)	0.999 (0.40%)	0.999 (0.20%)	0.999 (0.00%)	1.000 (0.00%)	0.999 (0.00%)	1.000 (0.00%)
N = 400	1.000 (0.20%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)
N = 600	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)
N = 1000	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)	1.000 (0.00%)

The table reports TLI values and accompanying rejection rates. Figure 3 shows the TLI values for the selected sample sizes and degrees of freedom. Results indicated that adding paths to the model (decreasing the degrees of freedom) did not do much in terms of altering TLI values. Increasing sample size resulted in larger TLI values. The model rejection rate did increase to 16% for a model with one degrees of freedom and a sample size of 50, suggesting that we could reject a correctly specified model when using the TLI and a combination of a small sample size and such a small degrees of freedom.

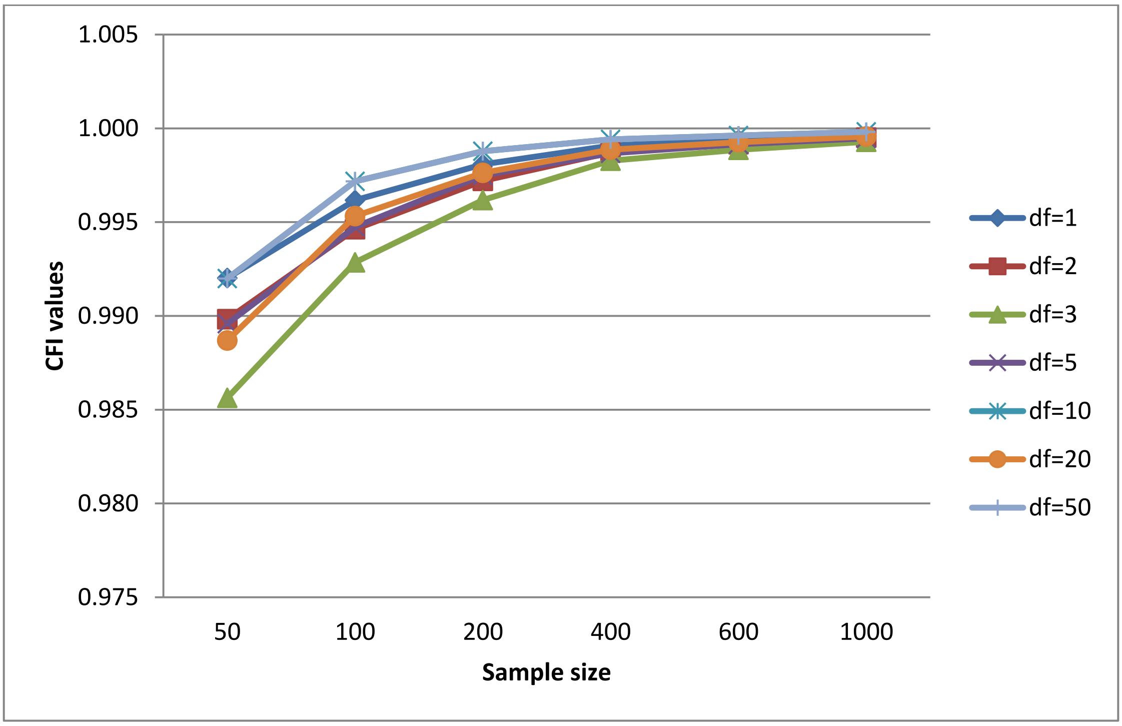


Figure 2. CFI values for selected degrees of freedom and sample size

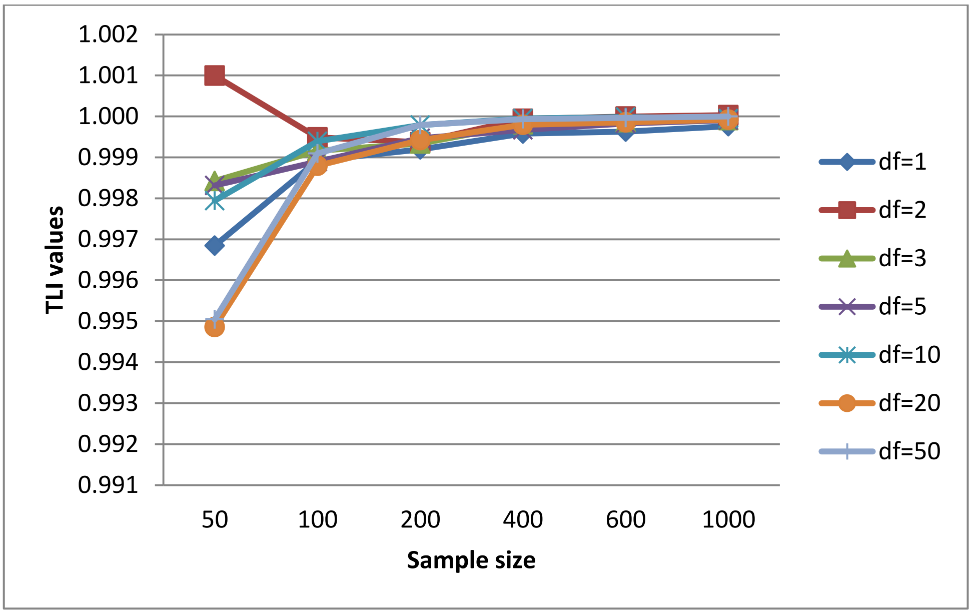


Figure 3. TLI values for selected degrees of freedom and sample size.

RMSEA. Results of the simulation for the various degrees of freedom and sample sizes are presented in Table 3. The table reports RMSEA values and accompanying rejection rates. Figure 4 shows the RMSEA values for the selected sample sizes and degrees of freedom.

Table 3. RMSEA values for selected degrees of freedom and sample size. Rejection rates are in parentheses.

Df	1	2	3	5	10	20	50
N = 50	0.053 (29.50%)	0.045 (30.10%)	0.048 (34.50%)	0.044 (34.20%)	0.041 (34.80%)	0.041 (35.60%)	0.041 (37.00%)
N = 100	0.037 (25.40%)	0.034 (26.30%)	0.034 (28.70%)	0.032 (26.60%)	0.027 (21.00%)	0.025 (16.00%)	0.021 (8.10%)
N = 200	0.027 (21.10%)	0.025 (20.10%)	0.025 (18.80%)	0.022 (14.60%)	0.019 (8.40%)	0.018 (2.70%)	0.014 (0.30%)
N = 400	0.019 (14.20%)	0.017 (8.00%)	0.016 (6.50%)	0.016 (4.30%)	0.013 (1.00%)	0.012 (0.00%)	0.009 (0.00%)
N = 600	0.016 (8.40%)	0.013 (5.00%)	0.014 (2.60%)	0.013 (0.90%)	0.010 (0.00%)	0.010 (0.00%)	0.008 (0.00%)
N = 1000	0.013 (4.90%)	0.010 (1.60%)	0.011 (0.70%)	0.010 (0.20%)	0.008 (0.00%)	0.007 (0.00%)	0.006 (0.00%)

Results indicated that adding paths to the model (decreasing the degrees of freedom) tend to result in larger RMSEA values. Increasing sample size resulted in smaller RMSEA values. It is also important to note that model rejection rates were high (greater than 30%) with a sample size of 50 regardless of the degrees of freedom. In addition, for an N of 100 and 200, models with smaller degrees of freedom had higher rejection rates. For example, the RMSEA exceeded the cutoff of .05 nearly 29% of the time with a sample size of 100 and degrees of freedom of 3. This is in line with findings by Kenny et al. (2014) indicating that researchers should proceed with caution when using the RMSEA with SEM models with small degrees of freedom and small sample sizes.

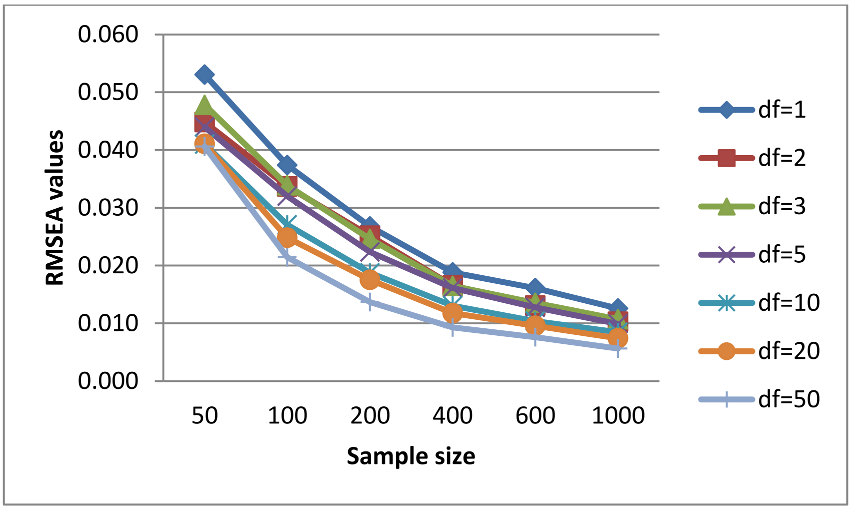


Figure 4. RMSEA values for selected degrees of freedom and sample size

SRMR. Results of the simulation for the various degrees of freedom and sample sizes are presented in Table 4. The table reports SRMR values and accompanying rejection rates. Figure 5 shows the SRMR values for the selected sample sizes and degrees of freedom.

Table 4. SRMR values for selected degrees of freedom and sample size.
Rejection rates are in parentheses.

df	1	2	3	5	10	20	50
N = 50	0.022 (0.20%)	0.047 (13.60%)	0.056 (18.50%)	0.053 (12.40%)	0.059 (15.80%)	0.064 (18.70%)	0.052 (3.10%)
N = 100	0.015 (0.00%)	0.034 (3.90%)	0.039 (4.30%)	0.037 (1.40%)	0.040 (1.20%)	0.043 (0.40%)	0.035 (0.00%)
N = 200	0.010 (0.00%)	0.024 (0.50%)	0.027 (0.20%)	0.026 (0.00%)	0.028 (0.00%)	0.031 (0.00%)	0.024 (0.00%)
N = 400	0.007 (0.00%)	0.017 (0.00%)	0.019 (0.00%)	0.018 (0.00%)	0.020 (0.00%)	0.022 (0.00%)	0.017 (0.00%)
N = 600	0.006 (0.00%)	0.013 (0.00%)	0.015 (0.00%)	0.015 (0.00%)	0.016 (0.00%)	0.018 (0.00%)	0.014 (0.00%)
N = 1000	0.005 (0.00%)	0.010 (0.00%)	0.012 (0.00%)	0.011 (0.00%)	0.013 (0.00%)	0.014 (0.00%)	0.011 (0.00%)

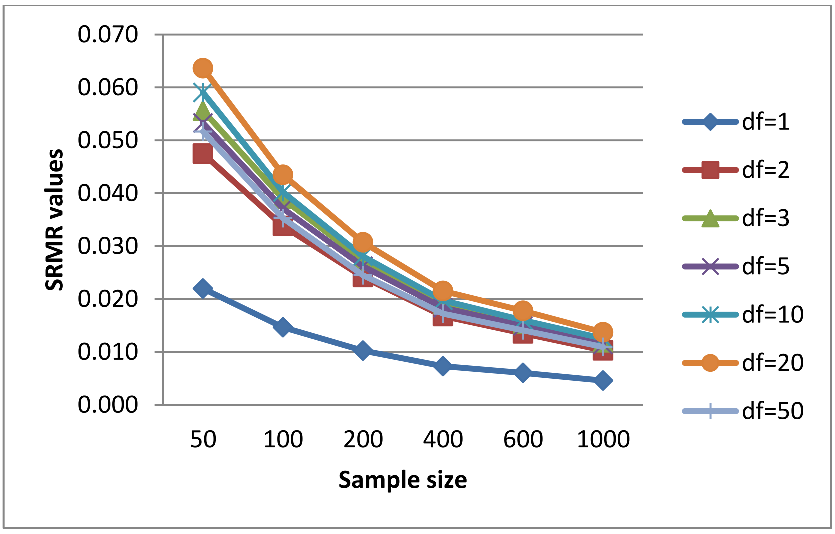


Figure 5. SRMR values for selected degrees of freedom and sample size.

Results indicated that adding paths to the model (decreasing the degrees of freedom) tend to result in smaller SRMR values. Increasing sample size resulted in smaller SRMR values. Once again, rejection rates tend to increase for a smaller sample size of 50 regardless of the degrees of freedom.

Conclusion

This is one of the first studies to examine the performance of the SRMR, RMSEA, CFI, and TLI with models with the small degrees of freedom typical of models found in math and science education. Kenny, Kaniskan & McCoach (2014) examined the RMSEA for correctly specified growth models with small degrees of freedom and we hoped to extend this line of research by examining the performance the RMSEA and additional fit indexes. We chose our four fit indexes because they are among the most widely reported in the SEM literature (Kline, 2010). Because we tested correctly specified models, we hoped that our fit indexes would not violate the cutoffs reported in the research and that rejection rates would be low. We manipulated sample size across various and small degrees of freedom and found that, in general, researchers should avoid sample sizes less than 100 when testing small degrees of freedom models. In fact, science and math education researchers should avoid reporting the RMSEA when sample sizes are smaller than 200, particularly when combined with small degrees of freedom. Small degrees of freedom do not tend to result in rejection of correctly specified models for the TLI, CFI, and SRMR, particularly if they tested using larger sample sizes.

References

1. Adedokun, O. A., Bessenbacher, A. B., Parker, L. C. Kirkham, L. L. and Burgess, W. D. **Research skills and STEM undergraduate research students' aspirations for research careers: Mediating effects of research self-efficacy,** Journal of Research in Science teaching, Vol. 50, No. 8, 2013, pp. 940-951
2. Akyol, G., Tekkaya, C., Sungur, S. and Traynor, A. **Modeling the interrelationships among pre-service science teachers' understanding and acceptance of evolution, their views on nature of science and self-efficacy beliefs regarding teaching evolution,** Journal of Science Teacher Education, Vol. 23, No. 8, 2012, pp. 937-957
3. Bailey, M., Taasoobshirazi, G. and Carr, M. **A Multivariate Model of Achievement in Geometry,** The Journal of Educational Research, Vol. 107, No. 6, 2014, pp. 440-461
4. Breivik, E., and Olsson, U. H. **Adding variables to improve fit: The effect of model size on fit assessment in LISREL,** Structural equation modeling: Present and future, 2001, pp. 169-194
5. Byars-Winston, A. M. and Fouad, N. A. **Math and science social cognitive variables in college students contributions of contextual factors in predicting goals,** Journal of Career Assessment, Vol. 16, No. 4, 2008, pp. 425-440
6. Chen, F., Curran, P. J. Bollen, K. A., Kirby, J. and Paxton, P. **An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models,** Sociological methods & research, Vol. 36, No. 4, 2008, pp. 462-494

7. Glynn, S. M., Taasoobshirazi, G. and Brickman, P. **Nonscience majors learning science: A theoretical model of motivation.** Journal of Research in Science Teaching, Vol. 44, No. 8, pp. 2007, pp. 1088-1107

8. Gottfried, A. E., Marcoulides, G. A., Gottfried, A. W. and Oliver, P. H. **A latent curve model of parental motivational practices and developmental decline in math and science academic intrinsic motivation.** Journal of Educational Psychology, Vol. 101, No. 3, 2009, p. 729

9. Ha, M., Haury, D. L. and Nehm, R. H. **Feeling of certainty: uncovering a missing link between knowledge and acceptance of evolution,** Journal of Research in Science Teaching, Vol. 49, No. 1, 2012, pp. 95-121

10. Hallquist, M. and Wiley, J. **Package 'MplusAutomation'.** 2013, retrieved from http://CRAN.Rproject.org/package=MplusAutomation

11. Hooper, D., Coughlan, J. and Mullen, M. **Structural equation modelling: Guidelines for determining model fit.** Articles, 2008, p. 2

12. Hu, L.-T. and Bentler, P. M. **Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives.** Structural equation modeling: a multidisciplinary journal, Vol. 6, No. 1, 1999, pp. 1-55

13. Kenny, D. A., Kaniskan, B. and McCoach, D. B. **The performance of RMSEA in models with small degrees of freedom.** Sociological Methods & Research, 2014, 0049124114543236.

14. Kenny, D. A. and McCoach, D. B. **Effect of the number of variables on measures of fit in structural equation modeling.** Structural equation modelling, Vol. 10, No. 3, 2003, pp. 333-351

15. Kirbulut, Z. D. **Modeling the Relationship between High School Students' Chemistry Self-Efficacy and Metacognitive Awareness.** International Journal of Environmental and Science Education, Vol. 9, No. 2, 2014, pp. 177-196

16. Kingir, S., Tas, Y., Gok, G. and Vural, S. S. **Relationships among constructivist learning environment perceptions, motivational beliefs, self-regulation and science achievement.** Research in Science & Technological Education, Vol. 31, No. 3, 2013, pp. 205-226

17. Kline, R. B. **Principles and Practice of Structural Equation Modeling, 3rd edn Guilford Press.** New York, USA, 2010

18. Mattern, N., and Schau, C. **Gender differences in science attitude-achievement relationships over time among white middle-school students.** Journal of Research in Science Teaching, Vol. 39, No. 4, 2002, pp. 324-340

19. Muthén, B. O. **Mplus technical appendices.** Los Angeles, CA: Muthén & Muthén,1998

20. Stevens, T., Olivarez, A., Lan, W. Y. and Tallent-Runnels, M. K. **Role of mathematics self-efficacy and motivation in mathematics performance across ethnicity.** The Journal of Educational Research, Vol. 97, No. 4, 2004, pp. 208-222

21. Worthington, R. L., and Whittaker, T. A. **Scale development research a content analysis and recommendations for best practices.** The Counseling Psychologist, Vol. 34, No. 6, 2006, pp. 806-838

[1] Gita Taasoobshirazi is an Associate Professor of Quantitative and Mixed Methods Research Methodologies at the University of Cincinnati. Her work involves examining and advancing the use of quantitative methods in education.

[2] Shanshan Wang is a doctoral student in Quantitative and Mixed Methods Research Methodologies at the University of Cincinnati. Her work involves understanding and advancing quantitative research methods in the social sciences.

HEURISTIC DECISION MAKING UTILIZING COMPLETE TOURNAMENTS[1]

Johan KOK[2]
Business Community Safety & Development Department
City of Tshwane, Republic of South Africa

E-mail: kokkiek2@tshwane.gov.za

R. JAYASREE
Research Scholar, Department of Mathematics,
Nirmala College for Women Coimbatore, India

E-mail: sree.ramachandran14@gmail.com

Abstract

The paper presents the newly developed Abarnica heuristic ranking method applied to fuzzy-bias decision making with the central technique derived from complete tournaments or derivatives thereof. An useful recursive proposition called, Akhil's proposition together with a challenging conjecture is presented. The challenge to derive an efficient algorithm to apply the Abarnica heuristic method which appears to be very efficient with manual applications remains open. The authors advocate that the main advantage of the Abarnica heuristic ranking method is that strong bias are analytically mitigated in fuzzy-bias decision making applications which require ranking.

Keywords: *Complete tournament, Perron-Frobenius theorem, Abarnica heuristic, corrupt driving testing*

1. Introduction

For general notation and concepts in graphs and digraphs see [1, 2, 3]. Unless mentioned otherwise all graphs are simple, connected and directed graphs (digraphs). Furthermore, if the context is clear the terms *vertex* and *player* will be used interchangeably.

A directed complete graph of order $n \geq 1$ with vertex set $V(G) = \{v_1, v_2, ..., v_n\}$ can be considered to be complete tournament, denoted by T_n. The understanding of a complete tournament is that all distinct pairs of vertices are players in a match and the arc (v_i, v_j)

indicates that player v_i beat player v_j. Unless mentioned otherwise a match between a distinct pair of players does not allow a draw. In other words it is assumed that an arc has distinct orientation. The assumption can easily be relaxed to analyze bi-orientations as well. Figure 1 depicts a complete tournament, T_6

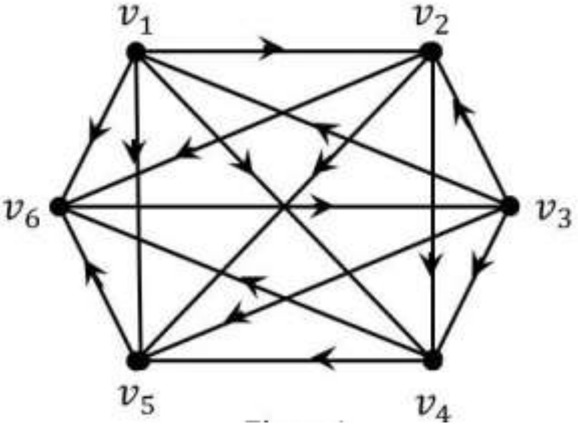


Figure 1.

If a vertex $v_i \in V(T_n)$ exists with out-degree $d^+(v_i) = n-1$, vertex v_i is the outright winner of the tournament. It is easy to see only one such vertex can exist. Also, if a vertex $v_j \in V(T_n)$ exists with in-degree, $d^-(v_j) = 0$, vertex v_j is the outright loser of the tournament. Only one such vertex can exist. For further analysis we shall only consider the subgraph $T - \{v_i, v_j\} = T_{n-2}^{'}$. The aforesaid is called the *reduction rule*. The reduction rule is applied iteratively until no further reduction is possible. Therefore, the final graph is either the empty graph hence, $V(T_0^{'''\cdots'}) = \phi$, if *n* is even, or $T_1^{'''\cdots'} = K_1$ if *n* is odd, or $T_k, k \geq 3$ results. Clearly for the first two outcomes the corresponding *trivial ranking* of the tournaments have been derived.

2. Important Graph Theoretical Results of Complete Tournaments

Unless mentioned otherwise, a resultant complete tournament say, $T_n, n \geq 3$ will be assumed. Figure 1 depicts a complete tournament that cannot be reduced further. It is assumed that the reader is familiar with the concept of a directed path and distance between vertices v_i and v_j. A *king vertex* v_i is a vertex for which the maximum directed distance between v_i and v_j, $\forall j \neq i$ is 2. For a directed path between vertices v_i and v_j of length 2 it is said that v_i gained a *virtual win* over v_j. Note that the aforesaid virtual win is possible despite a direct win of v_j over v_i.

Lemma 2.1 Without applying the reduction rule, a complete tournament T_3 has either one or three king vertices.

Proof: Up to isomorphism the only complete tournaments T_3 that exist are

$T_3 = (V(T_3), A(T_3)) = (\{v_1, v_2, v_3\}, \{(v_1, v_2), (v_1, v_3), (v_2, v_3)\})$ or

$T_3 = (V(T_3), A(T_3)) = (\{v_1, v_2, v_3\}, \{(v_1, v_2), (v_2, v_3), (v_3, v_1)\})$

Hence:

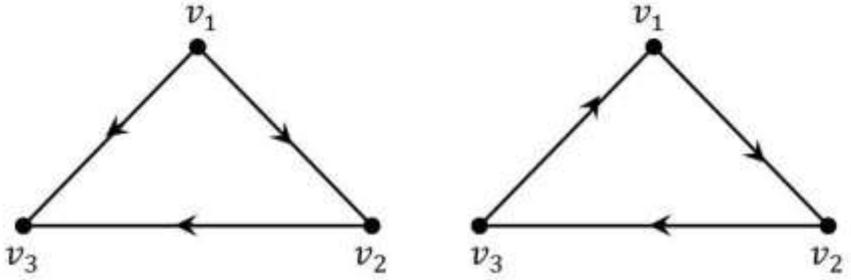


Figure 2.

Therefore, either only one king vertex, v_1 or three king vertices, v_1, v_2, v_3 exist.

Lemma 2.1 implies that for the case of having one king vertex the final ranking is $v_1, v_2\ v_3$

For the case of three king vertices the ranking requires at least one further match rule and/or at least one further heuristic criteria to resolve the ties. An example of heuristic criteria could be that player (vertex) v_1 displayed superior technique compares to that of players v_2, v_3 hence, the win of v_3 was a fluke. The heuristic criteria can be applied by say, a panel decision to break the tie between v_1 and v_3 or a re-match (match rule) is allowed.

A resultant complete tournament (reduction rule applied) is said to be diconnected if for any two distinct vertices, v_i, v_j there exists a directed path from v_i to v_j and from v_j to v_i.

Lemma 2.2 (Formalisation of observation in [1]) Up to isomorphism, there exists only one diconnected tournament of order 4.

Proof: Through exhaustive combinatorial re-orientation of arcs the figure below proves the result.

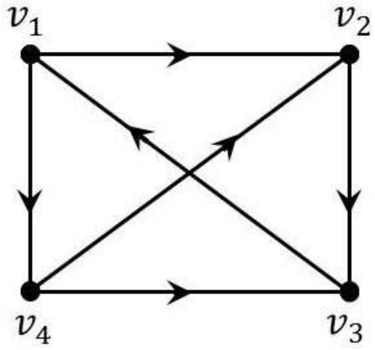


Figure 3.

Lemma 2.2 implies that all complete tournaments of order 4 other than the diconnected complete tournament can be trivially ranked by applying the reduction rule together with Lemma 1.1. From [1, pp. 185-188] it follows that all resultant complete tournaments on $n \geq 5$ vertices are disconnected and can be ranked by the Perron-Frobenius theorem.

2.1 The Abarnica Heuristic Ranking Method

The proposed Abarnica Heuristic Ranking Method[3] utilises the random selection of a sufficient number of primitive holes (see [4]) of a complete graph on $n \geq 4$ vertices to obtain only T_3 complete tournaments. These are then assembled to construct the complete tournament, T_n. This method has the advantage that different panels of judges or assessors can be utilised for the different T_3 tournaments when the rules are fuzzy-bias. Examples of such fuzzy-bias decision making tournaments are beauty contests, ranking political candidates, ranking music genres, ranking food brands, expert estimation of the impact speeds based on the crash damage to vehicles, agreeability within societies and alike. Hence, in decision making tournaments where the decision (game) rules are not a reliable approximation of objectivity.

Illustration 1. Consider the complete tournaments:

$$T_3^1 = (\{v_1, v_2, v_3\}, \{(v_1, v_2), (v_3, v_2), (v_3, v_1)\}),\ T_3^2 = (\{v_1, v_3, v_4\}, \{(v_3, v_1), (v_3, v_4), (v_4, v_1)\}),$$

$$T_3^3 = (\{v_1, v_3, v_5\}, \{(v_1, v_3), (v_5, v_3), (v_1, v_5)\}),\ T_3^4 = (\{v_1, v_5, v_6\}, \{(v_1, v_5), (v_5, v_6), (v_1, v_6)\}),$$

$$T_3^5 = (\{v_2, v_4, v_5\}, \{(v_2, v_4), (v_2, v_5), (v_4, v_5)\}),\ T_3^6 = (\{v_2, v_5, v_6\}, \{(v_2, v_5), (v_2, v_6), (v_5, v_6)\}),$$

$$T_3^7 = (\{v_3, v_5, v_6\}, \{(v_5, v_3), (v_6, v_5), (v_6, v_3)\}),\ T_3^2 = (\{v_4, v_5, v_6\}, \{(v_4, v_5), (v_4, v_6), (v_5, v_6)).$$

Since the number of panels and the stratification of the panel members can be arbitrary or well-structured (experts) we assume four panels namely P_1, P_2, P_3, P_4 provided the outcome above. Note that no panel could give absolute bias preference to any particular player (vertex). Also note that the tournaments were selected randomly. The only requirement to be met is that assembly into a singular tournament T_6 must result in a complete tournament with no tie in the orientation of an arc. In [4] it was shown that the number of primitive holes of a complete graph K_n is given by, $\binom{n}{3}$. It is easy to see that the assembly of the eight $T_3^i, i = 1,2,3,...,8$ tournaments results in the complete tournament depicted in figure 1. The observation that the complete tournament can be assembled by eight triangular tournaments compared to the twenty (20) distinct primitive holes found in K_6 signals great efficiency. In fact assembling is possible with only six (6) carefully selected primitive holes. Let $H(K_n)$ denote the minimum number of primitive holes of K_n to be oriented such that a complete tournament T_n can be assembled. In the next result the expression,

$\left\lfloor \frac{n}{2} \right\rfloor + 1$ is deliberately preferred above the simpler expression, $\left\lceil \frac{n}{2} \right\rceil$. The motivation is that the first expression most likely relates to the challenge to prove Conjecture 2.3.1.

Proposition 2.3 (Akhil's Proposition)[4] Any complete graph $K_n, n \geq 4$ has a minimum of $H(K_n) \leq H(K_{n-1}) + \left(\left\lfloor \frac{n-1}{2} \right\rfloor + 1 \right)$ primitive holes (K_3) (not necessary unique) for which the complete tournaments, $T_3^i, i = 1,2,3,...,H(K_n)$ on appropriate assembling, result in a complete tournament T_n (not necessary diconnected), provided that no arc orientation tie exists.

Proof: We begin by considering the complete tournament, T_3. For both of the only possible complete tournaments T_3^1 and T_3^2 it is clear that a mixed complete graph is obtained if vertex v_4 and edges v_1v_4, v_2v_4, v_3v_4 are added. Also, a minimum of two primitives holes say, on vertices v_1, v_2, v_4 and v_1, v_3, v_4 or v_1, v_2, v_4 and v_2, v_3, v_4 or v_1, v_3, v_4 and v_2, v_3, v_4 must be oriented. Thereafter, appropriate assembling any of the combinations will result in a complete tournament, T_4. Hence, the result $H(K_4) = 3 \leq H(K_3) + \left\lfloor \frac{3}{2} \right\rfloor + 1$ holds.

Assume the result holds for $n = k$ and consider any complete tournament, T_n.

Case 1: Let k be even and add vertex v_{k+1} and the edges, $v_1v_{k+1}, v_2v_{k+1}, v_3v_{k+1}, ..., v_kv_{k+1}$. Therefore, the maximum additional primitive holes to be orientated is, $\frac{k}{2}$. Since,

$H(K_{k+1}) \leq H(K_k) + \frac{k}{2} < H(K_k) + \left\lfloor \frac{k}{2} \right\rfloor + 1$, the result holds $\forall n$ is even.

Case 2: Let k be odd and add vertex v_{k+1} and the edges, $v_1v_{k+1}, v_2v_{k+1}, v_3v_{k+1}, ..., v_kv_{k+1}$. Therefore, the maximum additional primitive holes to be orientated is, $\left\lfloor \frac{k}{2} \right\rfloor + 1$. Since,

$H(K_{k+1}) \leq H(K_k) + \left\lfloor \frac{k}{2} \right\rfloor + 1$, the result holds $\forall n$ is odd.

Hence, through induction it follows that the results holds, $\forall n \geq 4$.

Conjecture 2.3.1 A minimum number of primitive holes, $H(K_n) \leq \left\lfloor \frac{\varepsilon(K_n)}{3} \right\rfloor + 1$ of a complete graph $K_n, n \geq 4$ can be orientated such that a complete tournament T_n can be assembled.

Consider figure 3 and note that vertices v_1, v_3, v_4 are king vertices. Also note that more than one distinct king path (directed path of length at most 2) may exist between vertices v_i and v_j.

Definition 2.1 The king index of a vertex v_i is the sum of the number of distinct king paths from v_i to v_j, $\forall j$. The king index is denoted, $k(v_i) \to V(T_n) - \{v_i\} = l, l \in N$.

Illustration 2. From figure 3 it follows that: $k(v_1) \to \{v_2, v_3, v_4\} = 2+2+1=5$,

$k(v_2) \to \{v_1, v_3, v_4\} = 1+1=2,$

$k(v_3) \to \{v_1, v_2, v_4\} = 1+1+1=3$ and

$k(v_4) \to \{v_1, v_2, v_3\} = 1+ +2 = 4$.

Since no king index ties exist the final ranking follows easily as: v_1, v_4, v_3, v_2.

Illustration 3. From figure 1 it follows that:

$k(v_1) \to \{v_2, v_3, v_4, v_5, v_6\} = 1+2+2+3+4=12,$

$k(v_2) \to \{v_1, v_3, v_4, v_5, v_6\} = 2+1+2+3=8,$

$k(v_3) \to \{v_1, v_2, v_4, v_5, v_6\} = 1+2+3+3+3=12,$

$k(v_4) \to \{v_1, v_2, v_3, v_5, v_6\} = 2+1+2=5,$

$k(v_5) \to \{v_1, v_2, v_3, v_4, v_6\} = 1+1+2+1+1=6$ and

$k(v_6) \to \{v_1, v_2, v_3, v_4, v_5\} = 1+1+1+1=4$.

The tie between vertices v_1, v_3 is resolved by noting that v_1 has king index summation term of 2 (total virtual wins) in respect of v_3 whilst v_3 has king index summation term of 1 (total virtual wins) in respect of v_1. Therefore, the derived ranking is, $v_1, v_3, v_2, v_5, v_4, v_6$. It is noted from [1] that the ranking derived is equal to the ranking derived by the Perron-Frobenius theorem.

Summarizing the Abarnica Heuristic Ranking Method

Step 1: Consider the complete graph $K_n, n \geq 4, n \in N$ and select at least $\left\lfloor \frac{\varepsilon(K_n)}{3} \right\rfloor + 1$ distinct primitive holes such that K_n can be assembled from these primitive holes.

Step 2: Choose a finite number of panels to decide on the orientation of the edges where an arc (v_i, v_j) will imply that vertex v_i is preferred, or is the winner over vertex v_j in terms of game rules or fuzzy-bias decision criteria.

Step 3: Assemble the complete tournament and apply the reduction rule to derive the preliminary ranking of all outright winners and losers. If the reduction rule results in a trivial ranking, exit. Else, go to Step 4.

Step 4: Determine the king index of all vertices of the resultant diconnected complete tournament and derive the complete ranking through the decreasing ordering of the king indices. If a tie occurs amongst king indices go to Step 5. Else, exit.

Step 5: For a king index tie between say, v_i and v_j, compare the virtual win score of v_i versus v_j as well as the virtual win score of v_j versus v_i. The highest virtual win score wins the ranking position. If a further tie occurs go to Step 6. Else, exit.

Step 6: If a virtual win score tie occurs between say v_i and v_j, the direct winner indicated by arc say, (v_j, v_i) is used to allow player v_j to win the ranking position.

3. Application to Driving License Testing within the South African Context

In the South African context the National Road Traffic Act, (Act 93 of 1996), provides for three clusters of driving licenses i.e. Code A1, A (motorbikes); B or EB (light motor vehicles) and Code C1, C, EC1 or EC (heavy motor vehicles). In section 3.2 of [5] a noticeable imbalance in the ratio of Code C1 license holders versus the population requiring that driving license code was observed. This observation raises the mysterious question for which a plausible answer should be researched. Why is it that so many decision makers and DLTC's and driving schools claim the existence of a high demand for Code C1 driving license testing for a vehicle category < 4, 1% of the vehicle population requiring that specific driving license code? The respected digital research company, Pondering Panda, released a survey in May 2013 in which it reliably reported that corrupt driving license testing is rife in South Africa. In particular, Code C1 is of interest because anecdotal evidence suggests strongly that Code C1 driving testing is the most corrupt driving testing code in the South African context. The aforesaid observation will be tested empirically by applying the Abarnica Heuristic Ranking Method to the complete tournament defined in the next section.

3.1. Complete Tournament Structure and Fuzzy-Bias Decision Making Rule

The complete tournament has players (vertices):

$$v_1 = \{A1, A\}, v_2 = \{B, EB\}, v_3 = \{C1\}, v_4 = \{C\}, v_5 = \{EC1\}, v_6 = \{EC\}.$$

The primitive holes to be orientated are on the set of vertices: $\{v_1, v_2, v_3\}$, $\{v_1, v_2, v_4\}$, $\{v_1, v_2, v_5\}$, $\{v_1, v_2, v_6\}$, $\{v_3, v_4, v_5\}$ $\{v_3, v_4, v_6\}$, $\{v_3, v_5, v_6\}$, $\{v_1, v_3, v_6\}$, $\{v_2, v_3, v_5\}$, respectively. The primitive holes were selected to ensure good empirical scrutiny of the Code C1 driving license. The selection allows for possible orientation ties between v_3 and all other players as well an orientation tie between all pairs amongst players v_2, v_3 and v_6.

The fuzzy-bias decision making rule is: Arc (v_i, v_j) = 1 implies that driving license testing in respect of Code(s) $\in v_j$ is least corrupt (or less prone to corrupt testing practices). Responses (decision making for orientation) will be captured in the table through 0 or 1 entries to mean that $(v_i, v_j) = 1 \Leftrightarrow (v_j, v_i) = 0$. Note that populating the complete table will serve as a statistical measure of consistency since a respondent may unintentionally or through subjective thinking map both the ordered pairs $(v_i, v_j) \to 1$ and $(v_j, v_i) \to 1$. As stated earlier, bi-orientations can be analyzed. Bi-orientation only nullifies a direct win but certainly contributes to the king index as well as the virtual win score as well. See the next illustrative table.

Table 1	v_i	v_j	v_k
v_i	-	0	1
v_j	1	-	1
v_k	0	0	-

3.2. Data Collection and Analysis

A total of 750 primitive holes were circulated for evaluation. A total of 665 responses were received hence, 1995 pairs of edges were evaluated (matched). Therefore, data collection had a response ratio of 88,67% which is considered satisfactory. Interesting to note is that 70,59% of non-respondents are from amongst practising examiners from the provinces, Western Cape, KwaZulu-Natal, Gauteng and Limpopo. The final tournament outcome is depicted in Table 2 below.

Table 2	A1/A	B/EB	C1	C	EC1	EC
A1/A	-	32	18	37	0	102
B/EB	88	-	62	61	35	20
C1	102	178	-	107	99	67
C	83	79	13	-	32	62
EC1	100	185	41	68	-	77
EC	13	100	33	78	23	-

From table 2 it is clear that a zero-row does not exist. It implies that no code category is viewed as corrupt free in respect of driver testing. It is observed that the Codes A1/A are viewed as corrupt free only in comparison of the heavy vehicle driving license code namely, EC1. Table 2 supports the anecdotal evidence that Code C1 is most prone to high levels of corrupt testing in that Code C1 has the highest average row score of 94,2. Table 3 defines the complete tournament and figure 5 depicts the complete tournament.

Table 3	A1/A	B/EB	C1	C	EC1	EC
A1/A	-	0	0	0	0	89
B/EB	56	-	0	0	0	0
C1	84	116	-	94	58	34
C	46	18	0	-	0	0

EC1	100	150	0	36	-	54
EC	0	80	0	16	0	-

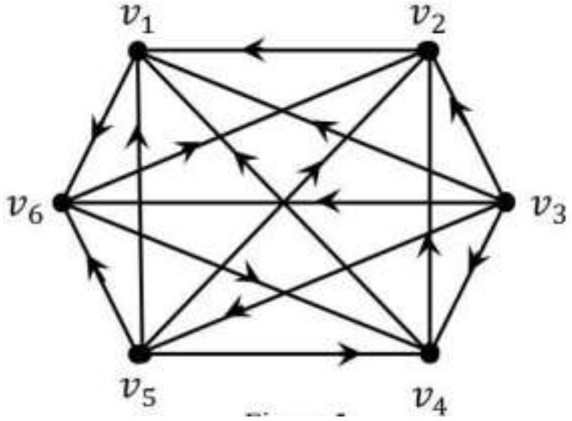


Figure 5.

Clearly vertex v_3 is the outright winner and upon reduction it follows that vertex v_5 is the second runner-up. Upon the second iterative reduction the resultant complete tournament on vertices v_1, v_2, v_4, v_6 is isomorphic to figure 3. Therefore, the final ranking is immediate i.e. $v_3, v_5, v_6, v_4, v_1, v_2$. The fact that vertex v_1 beats vertex v_2 comes somewhat as a surprise to the authors intuitive thinking. More so, since the average row score for Code A1/A is the lowest at 37,8. This result in itself requires some further investigation. A question that comes to mind is whether the growing popularity of superbikes and touring classics and/or the expanding scooter delivery and courier industry have resulted in a high demand for Codes A1 and A driving licenses.

4. Conclusion and Further Research

The application of the Abarnica Heuristic Ranking Method to the data indicates undoubtedly that the driving license testing for Code C1 can be considered as the most corrupt testing code within the South African context. The popularity of the Code C1 driving license probably lies in the fact that the driving test protocol is easier than that prescribed for a light motor vehicle. It is the authors considered view that most applicants who test for Code C1 actually do so with the intent to drive light motor vehicles. This view is currently the subject of research. However, mastering the easier driving competency might still be a challenge to many and therefore, the demand of corrupt testing and perhaps in many instances, no testing at all, to obtain a Code C1 is high. The aforesaid is then followed up by the Code EC1. The relationship between the first two driving license codes is that Code C1 is the lightest in the heavy vehicle cluster and Code EC1 is the lightest in the articulated heavy vehicle cluster. In terms of average row scores Codes B/EB and C performed in a close-tie with average row scores of 53,2 and 53,8 respectively. In [5] it was found that Code C is most likely becoming a redundant driving license code. It is suggested that the close-tie will be interesting to research further.

The power of the Abarnica Heuristic Ranking Method lies in the fact that the assessment of primitive holes mitigates strong bias or preference over a particular player (vertex) whilst the assembling of the primitive holes represents a complete tournament. The fact that Code A1/A beats Code B/EB and loses against Code C whilst Codes B/EB and C are in a close-tie illustrates the power of the Abarnica Heuristic Ranking Method to rank fuzzy-bias decision making applications.

It is clear that Akhil's proposition is a recursive result. It will be of interest to seek a closed formula and an efficient algorithm to find the minimum number of primitive holes $H(K_n)$. Any two primitive holes which do not share a common edge are said to be independent. Let the maximum number of independent holes of a graph G be denoted by, $\alpha^p(G)$. We propose an improved conjecture for further research.

Conjecture 4.1 For a complete graph K_n, $n \geq 3$, the minimum number of primitive holes of K_n to be oriented such that a complete tournament T_n can be assembled is given by

$$H(K_n) = \alpha^p(K_n) + \left\lceil \frac{\varepsilon(K_n) - 3\alpha^p(K_n)}{2} \right\rceil$$

References

1. Bondy, J.A. and Murty, U.S.R. **Graph Theory with Applications,** Macmillan Press, London, 1976
2. Chartrand, G. and Lesniak, L. **Graphs and Digraphs,** CRC Press, 2000
3. Harary, F. **Graph Theory,** Addison-Wesley Publishing Company, London, 1969
4. Kok, J. and Sudev, N.K. **A Study on Primitive Holes of Certain Graphs,** International Journal of Scientific & Engineering Research, Vol. 6, No. 3, March, 2015
5. Mathebula, S., Boadi-Kusi, S.M. and Kok, J. **Critical Review of Vision Fitness Testing within the South African Driving License Testing and Road Safety Context,** Journal of Applied Quantitative Methods, Vol. 11, No. 2, June 2016

[1] **Open access:** This paper is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution and reproduction in any medium, provided the original author(s) and the source are credited.

[2] Johan Kok (Ph.D. Applied Mathematics) is registered with the South African Council for Natural Scientific Professions in both the Mathematical Sciences and Physical Sciences categories. Johan has been endorsed by international peers as skilled in a wide range of combinatorica disciplines. His main research interests are in Graph Theory, Reconstruction of Vehicle Accidents and he has a keen interest in Mathematics Education. He also has a fondness for playing the most popular Africa drum called, the djembe.

[3] The first author dedicates this heuristic ranking method to young lady Abarnica, the niece to the second author.

[4] The first author dedicates this proposition to young lad Akhil, the nephew to the second author.

ON FINDING THE MOST COMPATIBLE BATTING AVERAGE

Prodip Kumar GAUR[1]
Post Graduate Student,
Department of Statistics, Assam University, Silchar, India

E-mail: prodip.786@rediffmail.com

Dibyojyoti BHATTACHARJEE[2]
Professor, Department of Statistics,
Assam University, Silchar, India

E-mail: djb.stat@gmail.com

Abstract

Batting average is the most commonly used measure of batting performance in cricket. It is defined as the total number of runs scored by the batsman divided by the number of innings in which the batsman was dismissed. Generally, the innings of a batsman comes to an end due to his dismissal, yet there are some cases in which the batsman may not get dismissed due to sudden termination of the batting innings of the team. The sudden termination may take place due to bad weather or victory or injury of the batsman or for running short of partners etc. In case, there are several not out innings in the career of a batsman, the batting average may get overestimated. To overcome this problem of over estimation, several authors proposed different modifications to the existing formula of batting average or defined new measures. Though each method expressed its advantages over the existing batting average, yet none of them are universally accepted as the most efficient replacement of the existing formula. This paper makes an attempt to study the existing solutions to the problem and then to evaluate the best or at least the most compatible alternative. For the purpose of quantification, data from the ICC Cricket World Cup played in Australia and New Zealand in 2015 is considered.

Keywords: *Data Mining in Sports; Performance Measurement; Cricket, Batting Average*

1. Introduction

Cricket is an outdoor game played with bat and ball in a specially prepared area in the center of circular field called a pitch. The game is played under certain rules and regulations between two teams of eleven players each. The teams take turn at batting and fielding. Each of such turn is called an innings. The aim of the fielding team is to dismiss all the batsmen of the batting team and/or to restrict the flow of runs. Presently, there are three versions of cricket being played at the international level: test cricket, one-day international cricket (ODI) and Twenty20 cricket (Saikia, Bhattacharjee & Radhakrishnan, 2016). While test match is an unlimited over game, ODI and Twenty20 are restricted over versions of cricket.

The ODI matches are of 50 overs per innings; and Twenty20, as the name indicates are of 20 overs duration only. ODI and Twenty20 format are called limited over format of cricket. In limited over cricket, the team which bats first sets a target for the opponent team to attain in the second innings.

In cricket each batsman try to score as much runs he can against the bowling attack of the fielding team. The bowlers on the other hand, with the help of the fielders try to restrict the batsman from scoring runs. Though both bowling and batting are the prime skills of the game of cricket yet in this work, we shall concentrate on a very common measure of batting performance viz. the batting average. The batting average is an index of the batting ability of a cricketer.

The batting average of a batsman in *'n'* innings (say) is defined as the total runs scored by the batsman in those innings divided by the number of complete innings. The phrase 'complete innings' means the innings in which the batsman was dismissed. If the batsman gets dismissed in all his innings then the batting average is as good as the arithmetic mean of the runs scored by the batsman in those *'n'* innings. However, if the batsman remains not out in some of the innings (antonymous to 'dismissed'), then the numerator remains same i.e. the runs scored by the batsman in those *'n'* innings but the denominator is only the innings in which the batsman was dismissed i.e. less than *n*. Thus, if there is at least one not out innings in the collection of *'n'* innings then the denominator is less than the number of terms in the numerator. This may overestimate the actual batting performance of a batsman, if measured through batting average. A hypothetical situation, in which the batsman is not dismissed in any of his innings, the batting average remains undefined. Many authors addressed this problem and defined different measures to compliment the issues concerning the batting average. Van Staden et al. (2009) gives a summary of all these methods. In this work, we try to explore these options and try to find out the best or the most compatible of the options.

Section 2 of the paper reviews the different types of performance measures in cricket and provides a brief introduction to all the extended batting averages defined by the different authors. The next section of the paper provides in details different formulae of all the extended batting averages. Section 4 is the methodology of the paper where the data source, process of comparison and relevant statistical methods are discussed. The result of the calculations is discussed in Section 5 and the last section concludes the work with some directions for future work.

2. Literature Review

Cricket is a data-rich sport. So different quantitative works by researches based on data generated from cricket are frequently encountered. Out of which a significant amount of work is concentrated towards performance measurement in cricket. Batting and bowling are two prime skills of the game. Thus, different traditional measures are used in cricket to quantify batting and bowling performance of cricketers. While batting average and strike rate are two very commonly used measures of batting performance, bowling average, bowling strike rate and economy rate are commonly used measured of bowling performance. In addition to these measures, several other authors have defined other innovative measures of quantifying batting and bowling skills of cricketers. Mention can be made of the Combine Bowling Rate by Lemmer (2002) and other measures of bowling performance by Beaudoin and Swartz (2003), Kimber and Hansford (1993) and Van Staden (2009). In case of innovative measures concerning batting performance mention can be made of Lemmer (2004), Barr and Kantor (2004), Croucher (2000), Basevi and Binoy (2007), Kimber and Hansford (1993). Brettenny (2010) reviews the different batting and bowling performance measures proposed by different authors.

Out of the traditional measures of quantifying batting performance in cricket the batting average is the most commonly used in all formats. The formula for which is given by,

$$\text{AV} = \frac{\text{Number of runs scored}}{\text{Number of complete innings}} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} x_i^*\right) \qquad (1)$$

Where $x_i; i = 1,2, \dots n$ denote the runs scored by a batsman in *n* completed innings and $x_i^*; i = n+1, n+2, \dots n+m$ denote the runs scored by a batsman in *m* not-out innings. The disadvantage of using this formula is that it can overestimate the batsman's batting average. Historically, the principle criterion used for comparing batsmen in the game of cricket has been the batting average, but unfortunately, when a batsman has a high proportion of not-out innings, the batsman's batting average will be inflated (Van Staden et al. 2009). The problem of the study looks beyond the works done by Van Staden et al. (2009).

To address this issue, several authors have suggested changes in the formula of batting averages with techniques ranging from the concept of survival analysis to Bayesian estimation. Some of them are Danaher (1989), Lemmer (2008a), Damodaran (2006), Maini and Narayanan (2007) etc. Though each of the methods is developed based on correct statistical logic yet there is no universal acceptance of any of these methods, as a solution to the problem of over estimation existing in (1).

Van Staden et al. (2009) analyzed and compared different methods which are designed to deal with the problem of inflated batting average due to the presence of a high proportion of not-out innings. From the work of Van Staden et al. (2009), one finds that none of the methods clearly outperforms all the other methods. His work only made an empirical comparison of ten different methods but cannot reach to a meaningful conclusion viz. the best method of computing the batting average. This provided us with a motivation to take up the problem.

3. Description of the different Batting Averages

The simplest solution for dealing with the problem of inflated batting average is to use the "real" AV instead of the conventional AV by dividing the number of runs scored in all innings by total number of innings,

$$AV_{real} = \frac{\text{Total runs scored}}{\text{Number of innings}} = \frac{1}{n+m}\left(\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} x_i^*\right) \qquad (2)$$

With AV_{real} the distinction between completed and not out innings is ignored, and, by doing so, the occurrence of inflated averages is completely eliminated (Howells, 2001).

Danaher (1989), proposed the product limit estimator (PLE) to estimate the batting average. The PLE is a non-parametric estimator originally designed by Kaplan and Meier (1958) for the use in life insurance and the actuarial field in general.

With the PLE, all not out batting scores are censored. Then,

$$PLE = \sum_{i=1}^{n} \Delta y_{i:n} \prod_{j=0}^{i-1} \left(1 - \frac{d_j}{c_j} \right) \tag{3}$$

Where $y_{i:n}; i = 1,2, \dots n$ denote the ranked distinct uncensored scores, $y_{0:n} = 0, \Delta y_{i:n} = y_{i:n} - y_{(i-1):n,} d_j$ is the number of uncensored scores equal to $y_{i:n}$ and c_j the number of censored and uncensored scores greater or equal to $y_{j:n}$. To ensure that the PLE is finite, the maximum score is uncensored, even if it is a not-out score.

Unfortunately the calculation of the PLE is extremely complex, so it is unlikely that the cricketing world shall favour it. Also, after each extra innings of a batsman, the PLE has to be recalculated completely. Furthermore, as pointed out by Danaher (1989), the PLE is insensitive when many of the high scores are not-out scores and hence censored.

Generally a batsman will always have PLE ≤ AV. However, it is interesting to note that the value of the PLE can be greater than that of AV. This can happen when a batsman's highest score is an outlier, that is, when the highest score is much larger than the second highest score and, in effect, the rest of the batsman's scores (Danaher 1989).

Lemmer (2008a) considered innovative estimators of the type

$$e_g = \frac{1}{n+m} \left(\sum_{i=1}^{n} x_i + f_g \sum_{i=n+1}^{n+m} x_i^* \right) \tag{4}$$

Where, the factor f_g is used to adjust the not out scores to obtain completed scores. The simplest estimator of this type is e_2 with, $f_2 = 2$, so that not out batting scores are doubled (Lemmer, 2008a).

$$e_2 = \frac{1}{n+m} \left(\sum_{i=1}^{n} x_i + 2 \sum_{i=n+1}^{n+m} x_i^* \right) \tag{5}$$

The justification for the choice of $f_2 = 2$ is that, if a batsman had a not out score and assuming that the batsman would be allowed to continue till he gets dismissed, then, on average, he could have been expected to double his score.

Lemmer (2008a), also considered many other possible factors, and found that e_6 with $f_6 = 2.2 - 0.01\overline{x}^*$, where $\overline{x}* = \frac{1}{m} \sum_{i=n+1}^{n+m} x_i^*$ is the average of the not out batting scores.

Thus we define e_6 as,

$$e_6 = \frac{1}{n+m}\left(\sum_{i=1}^{n} x_i + \left(2.2 - 0.01\bar{x}*\right)\right)\sum_{i=n+1}^{n+m} x_i * \qquad (6)$$

Lemmer (2008a) showed that e_2 and e_6 are closely related. But the calculation of e_6 is more complicated than that of e_2, accordingly he suggested that e_2 can be used for ease in calculations without much compromise with accuracy. Lemmer (2008b) recommended,

$$e_{26} = \frac{1}{2}\left(e_2 + e_6\right) \qquad (7)$$

Van Staden et al. (2009), defined another simpler measure like that of e_2 with an interesting modification. According to that method, the runs scored in the not out innings is either doubled or restricted to the highest score achieved by the batsman in the past tournament or career innings. Out of the two options, the minimum shall be considered for a not out innings. It is denoted by e_2^r.

Damodaran (2006), utilized a Bayesian approach to replace not out scores with conditional average scores. Consider the series of innings *t = 1, 2,…,n+m*, if the score in innings *t* is a complete score, x_t then we take $z_t = x_t$. If the score is a not out score, x_t^*, then this score is replaced by

$$z_t = \text{int}\left\{\frac{\sum_{l=1}^{t-1} z_l I\left(z_l\right)}{\sum_{l=1}^{t-1} I\left(z_l\right)}\right\} \qquad (8)$$

where $Z_1, Z_2, \dots, Z_{t-1}$ are the series of completed and/or adjusted scores up to innings $t-1, I(Z_l) = 0\ if\ x_i^* \geq Z_l and\ I(Z_l) = 1\ if\ x_i^* \leq Z_l$

The estimator for the average is then given by,

$$AV_{Bayesian} = \frac{1}{n+m}\sum_{t=1}^{n+m} Z_t \qquad (9)$$

Maini and Narayanan (2007) proposed a method based upon exposure-to-risk. Let

$$\bar{b} = \frac{\text{Number of balls faced}}{\text{Number of innings}} = \frac{1}{n+m}\left(\sum_{i=1}^{n} b_i + \sum_{i=n+1}^{n+m} b_i^*\right) \qquad (10)$$

Be the average number of balls faced by a batsman in his $m+n$ innings and let $r_1, r_2, \dots, r_n$ and $r_{n+1}^*, r_{n+2}^*, \dots, r_{n+m}^*$ denote the batsman's exposure in *n* completed innings and *m* not out innings respectively. If the score in innings i is a completed score, $r_i = 1.$ In effect the exposure is one for all completed innings. If the score is a not out score and $b_i^* < \bar{b},$ then, $r_i^* = \frac{b_i^*}{\bar{b}}$, else $r_i^* = 1$. The average is then calculated by

$$AV_{\exp osure} = \frac{\sum_{i=1}^{n} x_i + \sum_{i=n+1}^{n+m} x_i^*}{\sum_{i=1}^{n} r_i + \sum_{i=n+1}^{n+m} r_i^*} \tag{11}$$

Van Staden et al. (2009) pointed out two issues with the $AV_{exposure}$ - first, the number of balls faced by a batsman in a not-out innings is compared to the average number of balls faced over the whole tournament or career of this batsman. Thus, the exposure calculated for a not-out innings depends on past and future batting performances, which is not logical. Surely only past batting performances should be used. Further, the exposure for each past not-out innings must be recalculated each time the batsman bats again. So an immediate advantage of only using past batting performances will be that the exposure for past not-out innings need not be recalculated after each additional innings. The second concern has to do with the calculation of the average number of balls faced. Accordingly, Van Staden et al. (2009) suggested that a batsman should benefit from surviving the opposition's bowling attack by comparing the number of balls faced in a not-out innings to the survival rate instead of the average number of balls. Applying both the adjustments to the exposure-to-risk method, if a batting score is a not-out score and $b_i^* < SV_i$ where SV_i , is the survival rate for the batsman for all innings up to and including innings *i*, then $r_i^* = \frac{b_i^*}{SV_i}$, else $r_i^* = 1$. Accordingly, Van Staden et al. (2009) denoted the average based upon our adjusted exposure-to-risk method by $AV_{survival}$ to distinguish it from exposure AV_{exposure}.

4. Methodology

To compare the different averages discussed so far we need to apply it to some live data and compute the different averages. The conformity between the different methods shall be checked by the Kendall's coefficient of concordance and then in case of non-conformity sensitivity analysis shall be performed to find out the average that has maximum compatibility with the other averages. Detailed explanation of data source and the methodologies are explained in the subsequent subsections.

4.1. Data Source and Training Sample

For computing the batting averages using different methods and then for further relevant computations to reach the objective of the study we need a real data set. For the dataset, the matches played in the 2015 World Cup in Australia and New Zealand, is considered. The world cup of 50-overs a side saw 49 matches in the tournament. The necessary data from those matches are collected from the website www.espncricinfo.com. For the purpose of the study, the batsmen who satisfy the following criteria are considered in the trial sample for the computation of batting averages using the different methods:

- The batsman who has played at least 5 innings in the entire tournament
- The batsman who was not-out in at least one innings in the entire tournament
- The batsman who has faced at least 200 balls in the entire tournament

4.2. Computation of Averages

Following the restrictions as in the previous section, 20 batsmen qualified for the training sample, details of which are provided in Appendix I. Based on the different methods discussed above the computation are done and the averages along with the ranks are summarized in Table 1.

4.3. Kendall's Coefficient of Concordance

Since the formula for computation, varies from each other so it is obvious that the computed average using different methods will give different values even for the same batsman. However, the ranks of the batsmen shall not be much variant across the different methods, if based on the same data. Thus, once the averages of the batsmen are obtained using different methods, the batsmen shall be ranked based on each of the methods. Then considering each method of average as one of the rater, Kendall's coefficient of concordance shall be computed. Kendall's coefficient of concordance is a measure of agreement among raters and is defined as follows.

Assume there are *m* raters (here 10 different method of averages) rating *k* subjects (here 20 different batsmen) in rank order from 1 to *k*. Let $r_{ij} =$ the rating rater *j* gives to subject *i*. For each subject *i*, let $R_i = \sum_{j=1}^{m} r_{ij}$ let $\overline{R}$ be the mean of the R_i , and let *R* be the squared deviation, i.e.

$$R = \sum_{i=1}^{k} \left(R_i - \overline{R}\right)^2 \tag{12}$$

Now we define Kendall's *W* by

$$W = \frac{12R}{m^2\left(k^3 - k\right)} \tag{13}$$

It is also to be noted that the value of W always lies between 0 and 1 *i.e.* $0 \le W \le 1$. It is given that by the first property of Kendall's coefficient of concordance when $k \ge 5$ or *m* > 15, $m(k-1)W \sim \chi^2_{k-1}$. This rule can be used to test the null hypothesis that all the raters (averages) have ranked the subjects (batsmen) in a uniform manner.

4.4. Pareto Ordering for Compatibility

If the null hypothesis mentioned in the previous sub-section is rejected for the exercise on the current data set, then it means that the different methods of averaging have not ranked the batsmen in a uniform way but differently. In such a case, we can take the help of Pareto ordering to determine that average (set of ranks) which has the maximum compatibility with the other averages (set of rankings). Chakrabarty and Bhattacharjee (2012), can be consulted for detailed discussion of the Pareto ordering method. In brief, its working can be explained by the following way,

Let, the subscript *i* is an index attributed to identify the batsman. Since, there are 20 batsmen in the training sample so *i = 1, 2, …, 20* and the subscript *j* (or *k*) is an index attributed to the method of averaging. As, there are 10 method of averages discussed in the paper so *j* (or *k*) = *1, 2, …, 10* . Next, we define,

R_i^j = Rank of the *i*th batsman in the *j*th method of computing average

d_i^{jk} = Square of difference between ranks of the *i*th batsman for the *j*th and *k*th method of computing average = $\left(R_i^j - R_i^k\right)^2$

D^j = Sum of square of distance between ranks of the *j*th method of averaging with all other methods across all batsmen = $\sum_{k \neq j=1}^{10} \sum_{i=1}^{20} d_i^{jk}$ (14)

So, the compatibility score corresponding to the *j*th method of averaging is given by D^j as defined in (14). Lesser the compatibility score of a given method of average more is the compatibility of that average with a set of similar other method of averages.

5. Results and Discussion

Considering the data restriction mentioned above 20 batsmen got selected in the training sample. The batting average of all of them is calculated using the different method of averages and are placed in Table 1 below.

Table 1. Averages of the batsmen in the training sample under the different methods

Player name	**AV**	AV_{real}	**PLE**	e_2	e_6	e_{26}	e_2^r	$AV_{Bayesian}$	AV_{exposure}	$AV_{survival}$
SPD Smith	67(6)	57.43(5)	61.46(3)	65.43(8)	62.55(2)	63.99(5)	57.43(5)	57.43(5)	57.43(6)	59.04(5)
David Warner	49.29(13)	43.13(11)	48.058(9)	45.75(16)	45.72(10)	45.74(14)	43.13(11)	43.13(11)	48.14(9)	48.69(9)
Mahmudullah	73(3)	60.83(3)	56.67(7)	82.17(4)	59.13(4)	70.65(3)	60.83(3)	60.83(3)	60.83(3)	61.98(3)
IR Bell	52.4(11)	43.67(10)	44.75(13)	52.33(11)	49.56(7)	50.95(10)	43.67(10)	43.67(10)	43.73(11)	46.28(12)
MS Dhoni	59.25(9)	39.5(13)	44.5(14)	61.17(9)	34.63(14)	47.9(11)	39.5(13)	39.5(13)	39.5(13)	45.69(13)
Suresh Raina	56.8(10)	47.33(8)	41.33(15)	65.67(7)	49.17(8)	57.42(7)	47.33(9)	47.33(9)	47.33(10)	46.6(11)
Virat Kohli	50.83(12)	38.13(15)	46.42(11)	47.75(15)	37.23(12)	42.29(15)	38.13(15)	38.13(15)	39.23(14)	47.24(10)
Rohit Sharma	47.14(14)	41.25(12)	46.69(10)	48.38(14)	45.74(19)	47.06(12)	41.25(12)	41.25(12)	41.25(12)	42.97(14)
Ajinke Rahane	34.67(17)	29.71(17)	33.79(16)	33.43(17)	33.82(15)	34.12(17)	29.71(17)	29.71(17)	30.69(18)	32.22(18)
MJ Guptill	68.38(4)	60.78(4)	46.11(12)	87.11(3)	29.97(17)	58.84(6)	60.78(4)	60.78(4)	60.78(4)	55.78(7)
GD Elliot	44.29(15)	38.75(14)	60.98(4)	49.25(13)	42.53(11)	45.89(13)	38.75(14)	38.75(14)	38.75(15)	35.54(15)
KS Williamson	33.43(18)	26(19)	30.59(18)	32(18)	25.79(19)	28.9(19)	26(19)	26(19)	27.02(20)	31.59(19)
CJ Anderson	33(19)	28.88(18)	32.67(17)	29.75(19)	29.86(18)	29.81(18)	28.88(18)	28.88(18)	31.64(17)	32.34(17)
LRPL Taylor	31.57(20)	24.56(20)	29.7(19)	27.78(20)	24.64(20)	26.21(20)	24.56(20)	24.56(20)	27.2(19)	27.56(20)
AB de Villiers	96.4(2)	68.86(2)	75.67(2)	101.29(2)	53.7(6)	77.49(2)	78.14(1)	78.321(1)	69.42(2)	78.44(2)
DA Miller	64.8(7)	46.29(9)	50.95(8)	72.57(5)	36.82(13)	54.7(9)	52.86(8)	53.07(8)	48.42(8)	51.44(8)
F du Plessis	63.33(8)	54.29(7)	59.83(5)	57.29(10)	57.26(5)	57.27(8)	54.29(7)	54.29(7)	58.61(5)	59.72(4)
KC Sangakarra	108.2(1)	77.29(1)	101.63(1)	109(1)	62.86(1)	85.93(1)	77.29(2)	77.29(2)	77.29(1)	90.13(1)
MN Sammuels	38.33(16)	32.86(16)	19.29(20)	51.86(12)	30.39(16)	41.12(16)	32.86(16)	32.86(16)	32.86(16)	34.71(16)
SC Williams	67.8(5)	56.5(6)	59.82(6)	69.17(6)	62.07(3)	65.62(4)	56.5(6)	56.5(6)	56.5(7)	58.11(6)

In Table 1, the name of the batsman appears along the row heads and the method of average along the column heads. The number in the cell indicates the average of the batsman appearing in the row head using the method of average indicated by the column head. The numbers in parenthesis in each of the cells shows the rank of the cricketer depending on the method of averaging as given in the column head. It may be seen that KC Sangakara is ranked first in most of the method eight out of ten and LRPL Taylor is ranked last (20th) in eight out of ten methods.

Now, Kendall's coefficient of concordance is computed for the data set, with an aim to test the null hypothesis W = 0, which is an indication that there is agreement among the methods. The computation for the data set under consideration provides, W = 0.9034 with the *p*-value of the corresponding χ^2 statistic as 0.000 indicating that there is a clear disagreement between the different method of averages.

Next we perform Pareto ordering and compute the values of $D^{\,i}$ following (14) above. Table 2 provides the compatibility score of the different batting averages. Minimum the value of $D^{\,i}$ more compatible is the method of averaging. The table shows that, e_2^r and $AV_{Bayesian}$ has maximum compatibility with the other methods of averaging. AV_{real} acquires the next position in compatibility with very close compatibility score with e_2^r and $AV_{Bayesian}$. However, considering the simplicity of AV_{real} , one may consider it as the best method of computing batting average.

Table 2. Compatibility Score ($D^{\,i}$) of the different method of computing the batting average

AV	AV_{real}	*PLE*	e_2	e_6	e_{26}	e_2^r	$AV_{Bayesian}$	$AV_{exposure}$	$AV_{survival}$
1060	870	2278	1750	2988	952	868	868	1008	1090

6. Direction of Future Work and Conclusion

Of the different formats of cricket, this exercise is performed here only on one day international (ODI) cricket that too for a tournament only. In this exercise, we found that e_2^r and $AV_{Bayesian}$ has maximum compatibility compared to the other method of averages. In order to generalize the result, several such exercises shall be performed over different sets of one day international matches. The exercise can be extended to other formats of cricket like, Twenty20 and test cricket as well. This can enable the researcher to understand how the batting average shall be best defined depending on the format of cricket. However, what is statistically correct may not be accepted to cricket analysts and fans. The batting average needs to be well and simply defined, so that any cricket fan can easily compute the average. In this regard, in this exercise, AV_{real} even being in the second position shall earn more acceptance than others.

References

1. Barr, G. D. I and Kantor, B. S. **A criterion for comparing and selecting Batsmen in limited overs Cricket**, Journal of the Operational Research Society, Vol. 55, 2004, pp. 1266-1274
2. Basevi, T. and Binoy, G. **The worlds best Twenty20 players,** http://content-rsa.cricinfo.com/columns/content/story/311962html, 2007

3. Beaudoin, D. and Swartz, T. **The best Batsmen and Bowlers in One-day Cricket**, South African Statistical Journal, Vol. 37, 2003, pp. 203-222
4. Brettenny, W. **Integer optimisation for the selection of a fantasy league cricket team**, Unpublished Ph.D Dissertation, 2010
5. Chakrabarty, N. and Bhattacharjee, D. **Aggregation and weightage issues concerning composite index development: experience with digital divide in Asian countries**, Asia Pacific Journal of Library and Information Science, Vol. 2, 2012, pp. 24–37
6. Croucher, J. S. **Player rating in one-day cricket,** Proceedings of the Fifth Australian Conference on Mathematics and Computers. NSW, 2000, pp. 95-106.
7. Damodaran, U. **Stochastic dominance and analysis of ODI batting performance:the indian cricket team, 1989-2005**, Journal of Sports Science and Medicine,Vol. 5, No. 4, 2006, pp. 503-508
8. Danaher, P.J. **Estimating a cricketer's batting average using the product limit estimator**, New Zealand Statistician,Vol. 24, No. 1, 1989, pp. 2-5
9. Kaplan, E.L. and Meier, P. **Nonparametric estimation from incomplete observations,** Journal of American Statistical Association,Vol. 53, No. 282, 1958, pp. 457-481
10. Kimber, A. **A Graphical display for Comparing Bowlers in Cricket**, An International Journal for Teachers, Vol. 15, No. 3, 1993, pp. 84-86
11. Kimber, A. C. and Hansford, A.R. **A Statistical Analysis of Batting in Cricket**, Journal of the Royal Statistical Society , Vol. 156, No. 3, 1993, pp. 443-455
12. Lemmer, H H. **A Measure for the Batting Performance of Cricket Players**, South African Journal for Research in Sport, Physical Education and Recreation, Vol. 26, No. 1, 2004, pp. 55-64
13. Lemmer, H H. **The Combined Bowling Rate as a Measure of Bowling Performance in Cricket**, South African Journal of Research in Sport, Physical Education and Recreation, Vol. 24, No. 2, 2002, pp. 37-44
14. Lemmer, H. H. **An analysis of players' performances in the first cricket Twenty20 WOrld Cup series**, South African Journal for Research in Sport, Physical Education and Recreation,Vol. 30, No. 2, 2008b, pp. 71-77
15. Lemmer, H. H. **Measures of batting performance in a short series of cricket matches**, South African Statistical Journal, Vol. 42, No. 1, 2008a, pp. 65-87
16. Maini, S. and Narayanan, S. **The flaw in batting averages**, The Actuary, May-07, 2007, pp. 30-31
17. Saikia, H., Bhattacharjee, D. and Radhakrishnan, U. K. **A New Model for Player Selection in Cricket**, International Journal of Performance Analysis in Sports, Vol. 16, 2016, pp. 373-388
18. van Staden, P.J. **Comparison of cricketers bowling and batting performances using graphical displays**, Current Science, Vol. 96, No. 6, 2009, pp. 764-766
19. Van Staden, P.J., Meiring A.T., Steyn J.A., Fabris-Rotelli I.N. **Meaningful batting averages in cricket**, Annual Congress of the South African Statistical Association, November 2010, Potchefstroom, South Africa. Unpublished conference paper

Appendix 1. Innings wise performance of the batsmen in the training sample

Innings→ Batsman	1	2	3	4	5	6	7	8	9
SPD Smith	5(9)	4(11)	95(98)	72(88)	65(69)	105 (93)	56* (71)		
David Warner	22(18)	34 (42)	178 (133)	9(12)	21* (6)	24 (23)	12(7)	45 (46)	
Mahmadullah	23 (46)	28 (46)	62 (62)	103 (138)	128* (123)	21 (31)			
I R Bell	36(45)	8(17)	54(85)	49(54)	63(82)	52*(56)			
M S Dhoni	18(13)	18 (11)	45* (56)	85* (76)	6(11)	65(65)			
S R Raina	74(56)	6(5)	22(25)	110* (104)	65(57)	7(11)			
Virat Kohli	107 (126)	46 (60)	33* (41)	33(36)	44* (42)	38(48)	3(8)	1(13)	
Rohit Sharma	15(20)	0(6)	57*(55)	7(18)	64(66)	16(21)	137 (126)	34(48)	
Ajinka Rahane	0(1)	79 (60)	14(34)	33* (28)	19(24)	19(37)	44(68)		
MJ Guptill	49(62)	17 (14)	22(22)	11(14)	57(76)	105 (100)	237* (163)	34(38)	15 (34)
GD Elliot	29 (34)	29(31)	0(1)	19(28)	39(34)	27(11)	84(73)	83(82)	
KS Williamson	57(65)	38 (45)	9(22)	45(42)	33(45)	1(2)	33(35)	6(12)	12 (32)
CJ Anderson	75(46)	11 (16)	26 (42)	7(8)	39(26)	15(16)	58(57)	0(2)	
LRPL Taylor	14(28)	9(14)	5(5)	1(2)	24(41)	56(97)	42(61)	30 (39)	40 (72)
AB devilliers	25(36)	30 (38)	162* (66)	24(9)	77(58)	99(82)	65* (45)		
DA Miller	138*(2)	22(23)	20(16)	46*(23)	0(13)	49(48)	49(18)		
F duplessis	24(32)	55 (71)	62(70)	109 (109)	27(29)	21* (31)	82 (107)		
K Sangakarra	39(38)	7* (13)	105* (76)	117 (86)	104 (107)	124 (95)	45(96)		
Marlon Sammuels	21(41)	38 (52)	133* (156)	0(9)	2(7)	9(18)	27(15)		
SC Williams	8(13)	76* (65)	76(61)	33(32)	96(83)	50(57)			

Source: http://www.espncricinfo.com/icc-cricket-world-cup 2015/engine/series/509587.html? view=records
Note: Not out innings is indicated by *. Figures in bracket indicate the number of balls faced.

[1] Prodip Kumar Gaur, is a fresh Post Graduate from the Department of Statistics, Assam University, Silchar, Assam, India. His specialization is in Operations Research and Industrial Statistics.

[2] Professor Dibyojyoti Bhattacharjee has acquired his Post Graduate in Statistics and M. Phil (Statistics) both from the University of Delhi. He completed his Ph.D on Statistical Graphics from Gauhati University.
He worked in institutes like Central Statistical Organization, New Delhi, G C College, Gauhati University, Assam University. Currently he is the Head of the Department of Statistics at Assam University.
He has written about 15 books in different sub-fields of Statistics and published several research papers in different national and international reputed journals.

BRAIN DRAIN - BRAIN GAIN, EVIDENCE FROM THE EUROPEAN UNION

Mihaela GRECU
PhD, University Assistant, Department of Statistics and Econometrics,
Bucharest University of Economic Studies, Romania

E-mail: mihaela_grecu24@yahoo.com

Emilia TITAN
PhD, University Professor, Department of Statistics and Econometrics,
Bucharest University of Economic Studies, Romania

E-mail: Emilia_titan@yahoo.com

Abstract

The migration of skills is one of the most important and complex socio-economic phenomena. The mobility between the European Union countries gains more and more attention in the specialty literature. The migration process has strong economic implications – people are attracted by the better living and working conditions within the destination countries. Besides the economic, social and political implications this phenomenon presents ethical and moral implications too.

The direction of the migration for highly skilled persons is mainly from developing countries to the developed countries. The migration process can imply a loss and a gain at the same time. The countries of origin will suffer a loss of highly skilled/ educated people, while the receiving countries will gain without making any investment.

Keywords: *brain drain; brain gain; migration; European Union*

1. Introduction

The exodus of talented students to study abroad is the result of two factors. Firstly, the quality of domestic education; and secondly, it is a way for extraordinarily talented students to gain recognition of their skills. Even if the education received abroad is tangential to their ultimate employment, students may still choose foreign study to signal their exceptional quality.[1]

For young people that choose to continue their studies outside the country (higher education), the decision to migrate can be based on their expectations for future benefits. They either want a professional accomplishment; or they wish that when they return to their country of origin, to earn higher wages. In this case the temporary migration phenomenon is present and the diploma obtained represents a strong signal and an advantage over the candidates competing for the same position in the labor market. Alternatively, they want to settle in another country in which case the higher education achieved can increase their chances to be employed accordingly with their level of education. So, no matter if the labor mobility of the higher educated people is temporary or it is permanent, it attracts a series of advantages and disadvantages.

In terms of benefits from migration of highly skilled people (and not only) for the country of origin, here are some of the advantages: the labor market is stabilized, the unemployment rate decreased, the remittances are send in the country of origin which means an increase of the GDP for those countries. Generally, for the country of origin the disadvantages of the migration of the highly skilled people exceed the advantages.

As disadvantages we have: the quality and the number of higher education graduates that work in the country of origin decreases (the demand of highly skilled people falls) – phenomenon that lowers the chances for the country of origin to progress on medium and long term; the country of origin will only loose as long as it invests in education and it cannot make use of the future benefits.

In order to reduce the number of highly skilled people that leave their country, the state and the companies must take action. The curriculum of higher education institutions and demand in the labor market must be closely related. The young people must be prepared in order to adapt to the permanent changes that occur in the labor market – in this way the higher education institutes will have to review their curricula to the requirements of the labor market, so the young people can face the competition and find a job in line with their level of education. Another method by which the state or companies can stop the migration is adopting some laws in order to convince the highly skilled people to remain in their country. An example can be the IT domain in Romania. As an incentive to minimise the migration, the people that work in this field are exempted from paying income taxes.

There are cases in which we are confronted with the brain drain phenomenon, but not with the brain gain phenomenon too. If the highly skilled people worked in their country of origin on a position that reflects their level of education, but in the destination country they have a job that does not require higher education, we can say that only the brain drain phenomenon is present.

For the destination countries the number of advantages is bigger than the number of disadvantages. The main advantage is that the destination country wins without making any effort or investment. Hence, it benefits from highly skilled people. In this case we are facing with ethical implications as well. On the opposite side, the main disadvantage is that as long as the labor demand will be satisfied from the migration then on the long term this phenomenon can lead to a decrease in quality of the internal workforce.

If we are facing temporary migration, we can signal the presence of the brain exchange phenomenon, instead of the brain drain – brain gain phenomenon. When the brain exchange phenomenon is present we may say that we have a win-win situation for both countries of origin and destination.

If a person chooses to migrate only "virtual" – due to the advanced technologies, we can signal the presence of brain exchange phenomenon. In this way, the persons that are in this situation are living with their family without having to leave the country. They also bring benefits to their country – they spend the salary here and not in the "virtual" host country.

Attracting and keeping the performing labor force on the national market represents a condition of competitiveness, of ensuring sustainable development at local and national level.[2]

The main purpose of this study is to discuss the principal factors that lead people with tertiary education to migrate. Based on statistical analysis, we will investigate the flow of the migration process in the European Union (EU) countries for the highly educated people and determine what are the countries classified as countries of origin and what are the countries classified as destination countries.

The paper findings can be considered a good starting point for a better overview of the accuracy of the system.

2. Literature review and general framework

Some researchers think that the migration in the European Union shouldn't be seen as a disadvantage from the country of origin's perspective and that the main loss is that highly educated people leave the European continent and migrate to another continent, for example from Europe to USA. They say that Europe has to be considered as a whole.[3]

From our perspective, taking Europe's countries as a whole is almost impossible. In Europe there are developed countries, developing countries and undeveloped countries. The loss of highly educated people would mean a huge disadvantage, a loss that would diminish the chances of the countries of origin to accede to a higher level of education and livelihood.

According to the authors of the article "Brain Drain and Brain Gain Migration in the European Union after enlargement", you must fulfill two conditions in order to be included in the category of highly educated people from the migration perspective. The first condition refers to the level of the education achieved. According to this parameter, the highly educated people are included in the highly skilled or highly trained category. The second condition is strictly related to the profession and refers to the job that is practiced in the country of destination. Even if at a first glance the difference between these two parameters seems elusive, in reality it is not so. A highly educated person that works in the destination country as a taxi driver is framed in terms of education as an educated person, but not from the professional point of view too.[4]

3. Research goal, methodology and data issues

In the first part of the next section we tried to create a general framework of the migration process. To attempt to clarify which are the countries for which it is important to keep their highly educated people and which are the countries for which it is important to attract the highly educated people from abroad.

We collected the data from the Global Competitiveness Report for the period, 2013-2014. The two indices of interest are "Country capacity to retain talent" and "Country capacity to attract talent". They are included in the 7th pillar "Labor market efficiency" and have the codes 7.08 and 7.09. In the previous report there was only one index "Brain Drain" with code 7.07.

In the second part of the next section we developed a logistic regression model. Based on the above two indices we created a binary variable by taking into account the rank of the country of each of the two indices. The statistical tool used to perform the logistic regression is SPSS. The countries included in the analysis are: Austria, Belgium, Bulgaria, Czech Republic, Denmark, Germany, Greece, Spain, France, Italy, Hungary, Netherlands, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, United Kingdom, Norway and Switzerland.

4. Empirical results

The first step in our analysis is to create an image of the country ranking taking into account the "Country capacity to retain talent" and "Country capacity to attract talent".

According to the first index "Country capacity to retain talent" we have Finland on the second place (out of 148), followed by Switzerland (third place) and Norway (fifth place). On the opposite side we have Slovakia (130th place), Romania (138th place) and Bulgaria (142nd place).

If we take a look at the "Country capacity to attract talent" index we have top three countries as follows: Switzerland (first place), United Kingdom (fourth place) and Norway (11th place). At the end of the ranking we have Greece (127th place), Romania (132nd place) and Bulgaria (144th place).

If we analyze these two indices in parallel we can conclude that we have almost the same ranking position for most of the countries. The exceptions are Greece and Finland – for these two countries the capacity to retain talent is higher than the capacity to attract talent. For Greece the difference is 41 places in the ranking and for Finland are 66 places in ranking.

For Belgium we have a difference of 20 places in ranking, the capacity to retain talent being higher than he capacity to attract talent. For Czech Republic and Portugal we have as well at least 20 places in ranking between the two indices, but for these countries we have the index "Country capacity to attract talent" higher than the index "Country capacity to retain talent". For Czech Republic we have 22 places and for Portugal we have 23 places in ranking.

Table 1. Country capacity to retain talent ranking & Country Capacity to attract talent ranking

Country	Country capacity to retain talent Rank	Country	Country capacity to attract talent Rank
Finland	2	Switzerland	1
Switzerland	3	United Kingdom	4
Norway	5	Norway	11
Germany	9	Netherlands	18
Sweden	10	Germany	20
United Kingdom	13	Sweden	25
Netherlands	14	Austria	30
Austria	23	France	44
Belgium	26	Belgium	46
Denmark	43	Denmark	52
France	57	Finland	68
Greece	86	Czech Republic	87
Slovenia	107	Portugal	88
Spain	108	Spain	102
Czech Republic	109	Hungary	115
Portugal	111	Slovakia	119
Italy	117	Slovenia	120
Poland	119	Poland	121
Hungary	126	Italy	126
Slovakia	130	Greece	127
Romania	138	Romania	132
Bulgaria	142	Bulgaria	144

Source: Global Competitiveness Report for 2013-2014

In the next part we performed a correlation between all the indices included in our analysis. The "Country capacity to retain talent" index is strongly positively correlated with the indices "Life expectancy, years" (0.922), "Quality of overall infrastructure" (0.654) and "Quality of the educational system" (0.729). The index "Country capacity to attract talent" is positively correlated with "Quality of the educational system" (0.324).

We have correlations between the rests of the indices as well. The index "Life expectancy, years" is strongly positively correlated with "Quality of the educational system" (0.781), the index "Quality of overall infrastructure" is positively correlated with "Quality of the educational system" (0.798).

In the next part we developed a logistic regression model. The logistic regression model is used when the dependent variable is binary or qualitative and the independent variables are a mix of quantitative and qualitative variables.

The general form of the *Logit* model is:

$$\ln\left(\frac{p}{1-p}\right)=\alpha+\beta x+e \quad (1)$$

- *p*, represents the probability that the event *y* to occur: *p*(*y* = 1)
- ODD = *p*/(1- *p*) it's called "odds ratio"
- ln(*p*/(1- *p*)) are logarithms of odds ratio or logit[5]

In our analysis the dependent variable is "retain_attract2" variable which is obtained from the following formula:

- if retain_talent_rank-attract_talent_rank <= 7 then retain_attract2 = 0
- if retain_talent_rank-attract_talent_rank > 7 then retain_attract2 = 1

The value 0 for the new variable means "Attract" and value 1 means "Retain". The dependent variables are "life_expectancy" and "quality_education". Below we have our first output of this analysis – the "Classification Table".

Table 2. Classification Table

Observed			Predicted		
			retain_attract2		Percentage Correct
			Attract	Retain	
Step 0	retain_attract2	Attract	16	0	100.0
		Retain	6	0	.0
	Overall Percentage				72.7

Source: Author's work

Assuming that every country included in our analysis is an "Attract" country we get 72.7% classification accuracy. The model is also testing the hypothesis if the 6 of "Retain" and 16 of "Attract" countries are actually significant one from each other. If we have had the number of "Attract" equal with the number of "Retain" countries we would have an equal probability of being "Attract" and "Retain" countries. The next output tests that as a hypothesis.

We are rejecting the null hypothesis as we have the Sig.= 0.040 < 0.05, that there is an equal number of countries in each of the two groups. The odds ratio is calculated by dividing the number of "Retain" countries at the number of "Attract" countries. We have around 62.5% (1-0.375) chance that a country will not be an "Attract" country.

Table 3. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.981	.479	4.198	1	.040	.375

Source: Author's work

The last output from the first block is the "Variables not in the Equation". We can notice that if we take each independent variable separately, they are not statistically significant for this analysis, but if we take them together they are significant (Sig.= 0.045 < 0.05).

Table 4. Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	life_expectancy	2.599	1	.107
		quality_education	.006	1	.941
	Overall Statistics		6.199	2	.045

Source: Author's work

The next output if from block 1 and it's called "Omnibus tests of Model Coefficients". This output shows us the Chi-square and the Sig. values and tells us if you have at least some predictive capacity in the regression equation. Due to the fact that all the values for Sig. are significant we can assume that the independent variables are good predictors.

Table 5. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	6.692	2	.035
	Block	6.692	2	.035
	Model	6.692	2	.035

Source: Author's work

The output below shows us the predictive capacity of the model. This output is common to a lot of statistical analysis. We have the "-2 Log likelihood" which is similar to the "Chi-square", the "Cox & Snell R Square" and the "Nagelkerke R Square" values. The difference between the last two indices is the range; the "Cox & Snell R Square" index has a maximum value of 0.75, while the "Nagelkerke R Square" index has a maximum value of 1, so the last index will always have a larger value compared to the second index.

The most important index is the last one, being similar with the R Square from the linear regression (the calculation being different). We have 38% of the variance of the dependent variable is predicted by the independent variables.

Table 6. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	19.090	.262	.380

Source: Author's work

The "Hosmer and Lemeshow Test" gives us an idea of how good the model is. This time we want to have the Sig. value greater than 0.05, which is the case in our analysis (Sig.= 0.742) - so we can affirm that we have a good and significant model.

Table 7. Hosmer and Lemenshow Test

Step	Chi-square	df	Sig.
1	5.143	8	.742

Source: Author's work

Table 8. Contingency Table for Hosmer and Lemenshow Test

		retain_attract2 = Attract		retain_attract2 = Retain		
		Observed	Expected	Observed	Expected	Total
Step 1	1	2	1.941	0	.059	2
	2	2	1.924	0	.076	2
	3	2	1.847	0	.153	2
	4	2	1.822	0	.178	2
	5	1	1.792	1	.208	2
	6	2	1.697	0	.303	2
	7	2	1.486	0	.514	2
	8	1	1.236	1	.764	2
	9	1	.955	1	1.045	2
	10	1	1.301	3	2.699	4

Source: Author's work

The "Contingency Table for Hosmer and Lemenshow Test" output tells us how well the model is predicting certain outcomes. The main interest is for "Retain" – this will predict what country will be a "retain" or a "attract" one. If we look at the last step we have the observed number of subject value equal to 3 and our model predicted about 2.7. The closer these two values are, the better the model is.

In the "Classification table" we appreciate how good our model was in predicting the outcome. It is said that if the model is able to predict at least 65% of the categories it is a very good model.[6] Our model was able to predict 81.8% of the categories. Almost 82% of the outcomes were correctly predicted by our model. We have a greater value than the one from the null hypothesis where we had 72.7% (Classification table).

Table 9. Classification Table

Observed			Predicted		
			retain_attract2		
			Attract	Retain	Percentage Correct
Step 1	retain_attract2	Attract	14	2	87.5
		Retain	2	4	66.7
	Overall Percentage				81.8

Source: Author's work

In the last output of the analysis we have the values of the coefficients for the equation and the odds ratio Exp(B) as well. If the Exp(B) is greater than 1 the more likely the country is to be a "retain" country. For example if a country has a high quality of education value there are about 5.794 times more likely to be a "retain" country.

Table 10. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	life_expectancy	-2.531	1.179	4.611	1	.032	.080
	quality_education	1.757	1.032	2.897	1	.089	5.794
	Constant	.059	3.550	.000	1	.987	1.061

Source: Author's work

Our equation for this model is:

$$\log(p/(1-p)) = 0.059+1.757*\text{quality\_education}-2.531*\text{life\_expectancy}$$

The coefficient (or parameter estimate) for the variable quality education is 1.757. This means that for a one-unit increase in “quality_education”, we expect a 1.757 increase in the log-odds of the dependent variable retain_attract2, holding all other independent variables constant. For every one-unit increase in “life_expectancy”, we expect a 2.531 decrease in the log-odds of “retain_attract2”, holding all other independent variables constant. The expected value of the log-odds of retain_attract2 when all of the predictor variables equal zero is 0.059.

5. Conclusion

Regarding the two indices “Country capacity to retain talent” and “Country capacity to attract talent” we have countries that are in the top of the ranking for both indices: Switzerland, Norway; countries that are ranked differently for each of the two indices: Finland (66 places difference), Greece (41 places difference) and countries that are at the bottom of the ranking for both indices: Bulgaria, Romania, Poland. The last three countries mentioned will always suffer because of the loss of highly educated people. People from these countries will migrate most probably to the countries included in the first group presented above.

The logistic regression conducted predicts a very accurate model, by using the two independent variables life expectancy and quality of education. These two variables explain around 38% of the variability of the model. The model is very sensitive to the increase of the quality of education index; this means that a small increase of this index will be highly observed in the predicted model.

References

1. Hilbe, M. J. **Logistic Regression Models,** CRC Press, Taylor and Francis Group, 2009
2. Hilbe, M. J. **Practical Guide to Logistic Regression,** CRC Press, Taylor and Francis Group, 2015
3. Hosmer, D. W. and Lemeshow, S. **Applied Logistic Regression,** Second Edition, John Wiley & Sons Inc, New York, United States, 2001
4. Kelo, M. and Wachter, B. **Brain Drain and Brain Gain Migration in the European Union after enlargement,** European conference Braingain – the instruments, The Hague, 2004
5. Knok, V. and Leland, H. **An Economic Model of the Brain Drain,** The American Economic Review, Vol. 72, No. 1, 1982
6. Vasile, V. **Economic and Social Inferences of the Highly Skilled Labor Force Migration,** Buletinul Universitatii Petrol – Gaze din Ploiesti, Vol. LVIII, No. 1, 2006
7. https://cristinaboboc.files.wordpress.com/2016/04/curs-8-9-2016-statistica-neparametrica1.pdf
8. http://www3.weforum.org/docs/WEF_GCR_Report_2011-12.pdf
9. http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2013-14.pdf

10. http://www.statisticssolutions.com/mlr/
11. http://www.ats.ucla.edu/stat/stata/dae/logit.htm

Appendix

Correlations

		Brain Drain Indicator	Country capacity to retain talent	Country capacity to attract talent	Life expectancy, years	Quality of overall infrastructure	Quality of the educational system
Brain Drain Indicator	Pearson Correlation	1	.945**	.158	.921**	.683**	.758**
	Sig. (2-tailed)		.000	.484	.000	.000	.000
	N	22	22	22	22	22	22
Country capacity to retain talent	Pearson Correlation	.945**	1	.110	.922**	.654**	.729**
	Sig. (2-tailed)	.000		.627	.000	.001	.000
	N	22	22	22	22	22	22
Country capacity to attract talent	Pearson Correlation	.158	.110	1	.102	.264	.324
	Sig. (2-tailed)	.484	.627		.653	.235	.141
	N	22	22	22	22	22	22
Life expectancy, years	Pearson Correlation	.921**	.922**	.102	1	.607**	.781**
	Sig. (2-tailed)	.000	.000	.653		.003	.000
	N	22	22	22	22	22	22
Quality of overall infrastructure	Pearson Correlation	.683**	.654**	.264	.607**	1	.798**
	Sig. (2-tailed)	.000	.001	.235	.003		.000
	N	22	22	22	22	22	22
Quality of the educational system	Pearson Correlation	.758**	.729**	.324	.781**	.798**	1
	Sig. (2-tailed)	.000	.000	.141	.000	.000	
	N	22	22	22	22	22	22

**. Correlation is significant at the 0.01 level (2-tailed).

[1] Knok, V. and Leland, H. **An Economic Model of the Brain Drain,** The American Economic Review, Vol. 72 No. 1 1982, p. 98

[2] Vasile, V. **Economic and Social Inferences of the Highly Skilled Labor Force Migration,** Buletinul Universitatii Petrol – Gaze din Ploiesti, Vol. LVIII, No. 1, 2006, p. 15

[3] Kelo, M. and Wachter, B. **Brain Drain and Brain Gain Migration in the European Union after enlargement,** European conference Braingain – the instruments, The Hague, 2004

[4] Idem, p. 17

[5] https://cristinaboboc.files.wordpress.com/2016/04/curs-8-9-2016-statistica-neparametrica1.pdf

[6] http://www.statisticssolutions.com/mlr/

THE COMPLETENESS AND ACCURACY OF INFORMATION ABOUT COELIAC DISEASE ON THE ROMANIAN WEBSITES

Valentin NADAŞAN[1]
PhD, Lecturer,
University of Medicine and Pharmacy, Targu Mureş, Romania

E-mail: vnadasan@gmail.com

Olesea MOLDOVAN
Master in Community and Clinical Nutrition,
University of Medicine and Pharmacy, Targu Mureş, Romania

Abstract

The internet has become an important source of health related information and a number of studies have shown that the quality thereof is, at best, problematic. Nevertheless, there are very few studies investigating the Romanian medical cyberspace. The goal of this study was to assess the completeness and accuracy of information about the coeliac disease on the Romanian websites directed to the general population. We evaluated a sample of 100 websites selected from the Google's first search results pages. The coverage of the topic was extremely deficient (the mean completeness score was 3.8 on a scale of 10), especially on sensitive issues such as the causes, treatment, and complications of the coeliac disease. On the other hand, the accuracy of the information was relatively good (mean accuracy score 7.2 on a scale of 10). With one exception, we found no statistically significant differences between the quality scores of the websites by their general characteristics.

Keywords: *coeliac disease; gluten intolerance; consumer health; information quality; Internet*

Introduction

The Internet has become a major source of health-related information available for most of the general population in the developed and developing countries (Eysenbach, 2000; Boyer, 2010). The latest surveys conducted in North America and Europe have shown that a proportion of 50 to 80% of the population have searched health-related information on the Internet (Wang et al, 2012; Seybert, 2011). A Romanian public survey reported that 96% of the Romanians were using the Internet to seek information and 15% of the respondents were looking for health-related information during their online searches (IRES, 2011). The poor quality online health information is an emerging public health concern as it exposes the unaware consumers to notable health risks by delaying imperative interventions, experimenting with ineffective or dangerous treatments, followed by aggravating of the disease or death (Weaver et al, 2009; Eng & Gustafson, 1999).

Although nutrition and nutritional diseases are listed among the most searched for topics among people of all ages, (Wang et al, 2012; Ettel et al, 2012), and coeliac disease and gluten intolerance seem to draw the attention of a wide group of population and the mass-media, the quality of information about these conditions on the Romanian Internet is virtually unknown.

The aim of the study was to answer the following research questions:

1. What are the general characteristics of the Romanian coeliac disease websites?
2. What is the quality of the information about coeliac disease, as far as completeness and accuracy are concerned?
3. Are there any significant differences regarding the quality of the information about the coeliac disease between the websites with different general characteristics?

2. Material and methods

The study was designed as an observational cross-sectional study. The sample included the first 100 websites listed on the Google search engine's results pages. The search was conducted during March-April 2016 using "boala celiacă" ("coeliac disease") as a query term. We limited the search to the Romanian websites by using the language specific search page (www.google.ro).

We included only those sites that covered the topic under investigation in at least 250 words in Romanian language and which were targeted to the general population. Sponsored links, discussion forums, video- or audio-only content, infected or unavailable sites and also sites that required registration were all excluded. When multiple pages or subdomains belonging to the same top level domain were listed as separate links on the search engine's results page, we treated them as one website.

We classified the websites by their general characteristics: type of ownership (private individual, foundation/association, educational or research institution, public or private health service provider, commercial society), main purpose (educational, commercial, networking), genre (thematic, online newspaper or journal, corporate website, online store, blog or personal website, other) and medical paradigm (conventional, alternative, mixed). The definition of each website category and the description of the assessment procedure were included in an assessment form that was made available to the person performing the evaluation.

The quality of the information about the coeliac disease was measured using two generally accepted quality criteria: completeness and accuracy. The content of each website was checked against a list of expected items that we developed from the evidence-based medical literature. This standard content list was also included in an assessment form along with comprehensive instructions for the evaluators. (The assessment forms are available upon request from the corresponding author.) Each website was rated for completeness and accuracy. The raw scores were eventually converted to a relative completeness score (rCS) and a relative accuracy score (rAS) ranging from 0 to 10, to facilitate comparison within the sample. The website content grading procedure and quality score calculations are described in detail in a previous work (Nădăşan et al, 2011).

We checked for statistical differences between the quality scores of the websites classified by their general characteristics with the nonparametric Kruskal-Wallis test. The level of statistical significance was set at 0.05. All statistical analyses were carried out using GraphPad InStat Demo 3.06.

Results

Descriptive data about the general characteristics of the coeliac disease websites included in the studied sample are presented in table 1.

Table 1. The general characteristics of the Romanian websites presenting information about the coeliac disease

Characteristics	Subcategories	N (%)
Ownership	Individual	10 (10)
	Foundation, association	17 (17)
	Educational institution	2 (2)
	Health service provider	10 (10)
	Commercial society	31 (31)
	Unidentifiable	30 (30
Purpose	Educational	83 (83)
	Commercial	15 (15)
	Networking	2 (2)
Genre	Thematic	27 (27)
	Online newspaper, journal	12 (12)
	Corporate website	24 (24)
	Online store	9 (9)
	Blog, personal website	18 (18)
	Other genre	10 (10)
Medical paradigm	Conventional	33 (33)
	Alternative	7 (7)
	Mixed	20 (20)
	Unidentifiable	40 (40)

The mean rCS for the whole sample was 3.8 points and the rAS was 7.2 points (on a scale ranging from 0 to 10). The mean rCS and rAS of the websites categorized by their general characteristics are shown in table 2.

Table 2. Mean rCS and rAS of the Romanian websites presenting information about the celiac disease by subcategory

Characteristic	Subcategory	Mean RCS (SD)	Kruskal-Wallis H (p-value)	Mean RAS (SD)	Kruskal-Wallis H (p-value)
Ownership	Individual	3.9 (2.1)	745 (0.3402)	7.9 (1.3)	5.297 (0.3807)
	Foundation, association	3.8 (2.0)		7.6 (1.2)	
	Educational	5.5 (0.7)		07.06.16	

Characteristic	Subcategory	Mean RCS (SD)	Kruskal-Wallis H (p-value)	Mean RAS (SD)	Kruskal-Wallis H (p-value)
	institution				
	Health service provider	3.4 (1.9)		7.0 (1.1)	
	Commercial society	3.7 (1.8)		7.3 (1.2)	
	Unidentifiable	4.5 (2.0)		7.4 (1.0)	
Purpose	Educational	3.9 (2.0)	0.8542 (0.6524)	7.4 (1.1)	0.9570 (0.6197)
	Commercial	4.1 (2.0)		7.3 (1.1)	
	Networking	5.0 (1.4)		7.0 (0.1)	
Genre	Thematic	4.9 (1.9)	5.269 (0.0988)	7.5 (1.1)	2.178 (0.8240)
	Online newspaper	3.7 (1.8)		7.6 (0.9)	
	Corporate website	3.7 (1.6)		7.3 (1.1)	
	Online store	3.5 (1.7)		7.5 (1.4)	
	Blog, personal website	3.8 (2.4)		7.5 (1.2)	
	Other genre	3.2 (1.8)		7.0 (1.1)	
Medical paradigm	Conventional	4.4 (1.9)	9.384 (0.0246)	7.7 (0.9)	3.896 (0.2729)
	Alternative	5.4 (1.5)		6.8 (1.2)	
	Mixed	3.7 (1.9)		7.3 (1.1)	
	Unidentifiable	3.5 (1.9)		7.4 (1.2)	

The mean rCSs and mean rASs of information calculated separately for the main subsections of the coeliac disease are represented in figure 1.

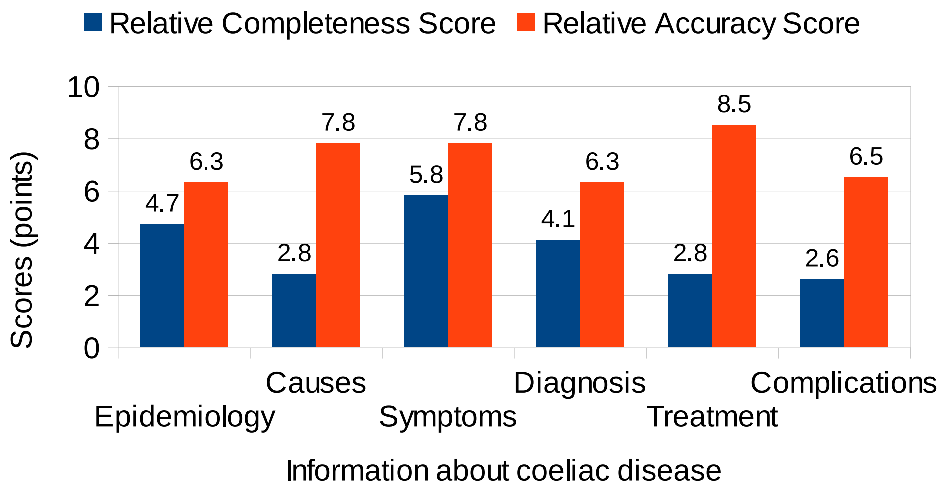


Figure 1. The mean relative completeness score and relative accuracy score of the information about coeliac disease on the Romanian websites by subsections

Discussions

To the best of our knowledge, this is the first study to assess the quality of the coeliac disease-related information on the Romanian websites.

Our findings show that most of the websites are owned by commercial entities and there is an obvious lack of involvement of the educational and research institutions. Considering that the coeliac disease represents an opportunity to merchandise gluten-free packaged foods, the prominent online presence of commercial companies among the websites disseminating information about this nutritional disease seems a natural consequence since a web page is one of the most cost-efficient promotional solution. While one would expect that educational or research institutions would produce and publish higher quality information, our study was not able to identify a significant difference between the quality of information related to coeliac disease available on the websites owned by commercial or non-commercial entities.

As far as the medical paradigm of the websites, the field seems to be about equally shared by those who embrace the conventional (33 websites) and the alternative or mixed approach (27 websites). The results of our assessment have shown a slightly higher mean completeness score for the alternative medicine websites but this observation needs to be taken with caution as the magnitude of the difference might be of rather statistical than practical significance. The explanation of the somewhat higher completeness scores observed among the alternative medicine websites in this study might be related to the fact that the primary therapy for ceoliac disease at this time consists of a gluten-free diet, which is probably perceived, unwarrantedly, as an alternative method of treatment and so, this type of websites tend to provide more detailed information on the topic. However, it should be noticed that a previous more comprehensive research on the quality of health-related Romanian websites have not reported such a superiority of the alternative medicine websites compared to the conventional medicine websites (Nădăşan, 2011).

The main findings of our study show that overall, the completeness of the information presented on the Romanian coeliac disease websites is extremely deficient (the mean rCS was 3.8 on a scale of 10) while the overall accuracy of the information is apparently good (the mean rAS was 7.2 on a scale of 10). A similar paper investigating the quality of information related to coeliac disease on English language websites reported that not more than 50% of the information provided at least half of the required information related to coeliac disease and 52% of the websites have reached a level of accuracy of 95% (McNally et al, 2012). The authors conclude that the lacking accuracy and transparency of the investigated English language websites about coeliac disease makes them unreliable for both medical professionals and general users.

Besides the quality assessment conducted on each website as a whole, we assessed also the quality of the information by each separate chapter of the coeliac disease. These additional analyses revealed that there are sensitive sections, especially the information regarding the treatment and complications of coeliac disease (rCS was 2.8 and 2.6 respectively) that are deficiently covered and thus represent a potential risk for the patients.

It should be underlined that the completeness and accuracy scores used in this study must not be interpreted independently because the completeness score measures exclusively the coverage of the topic while the accuracy score exclusively the correctness of the information without any reference to completeness. Therefore, websites with unacceptably low coverage of the topic can receive a high or very high accuracy rating if the information they do provide, is free from errors.

At a closer examination, we found that only three of the websites in the Romanian sample had both completeness and accuracy scores in the acceptable range (>7.5 to 10 points). This observation might have practical implications since the probability of finding exhaustive and simultaneously correct information about the topic on a single website is extremely low. Thus, the likelihood of inexperienced users who do not spend sufficient time to cross-check the information on several websites, to be exposed to incomplete and/or inaccurate information seems to be very high.

The main limitation of the study is inherently related to the Internet research. Given the fluidity and volatility of the online content, the replication of the study seems virtually impossible. Using different search engines or query terms is likely to significantly alter the elements and hierarchy of the sample and thus an upward or downward shift in the quality scores as well.

We attempted to minimize the subjectivity of the assessment by developing a detailed description of the assessment methodology. However, the results of our evaluation might be affected by the subjective nature of certain aspects of the rating procedure.

Although the scope of this study was limited to a narrow field of the Romanian medical cyberspace, the reported observations bring new knowledge to the general picture of the quality of health-related information on the Romanian Internet.

Conclusions

1. The coverage of the coeliac disease as a health topic on the Romanian websites was extremely deficient (mean completeness score 3.8 on a scale of 10).
2. The extreme lack of information on sensitive issues such as the causes, treatment, and complications of the coeliac disease should be a real concern.
3. The accuracy of the information about the coeliac disease on the Romanian websites was relatively good (mean accuracy score 7.2 on a scale of 10).
4. With one, probably irrelevant exception, we found no statistically significant differences between the quality scores of the websites by their general characteristics.

References

1. Boyer, C. **Education and consumer informatics,** Yearb Med Inform, 2010, pp. 72-74
2. Eng, T.R. and Gustafson, D.H. (editors) **Science Panel on Interactive Communication and Health. Wired for Health and Well-Being: the Emergence of Interactive Health Communication,** US Department of Health and Human Services, Washington, DC, 1999, accessed 04.05.2016, http://www.ehealthstrategies.com/files/eng_gustafson_1999.pdf
3. Ettel, G. 3rd, Nathanson, I., Ettel, D., Wilson, C. and Meola, P., **How do adolescents access health information? And do they ask their physicians?**, Perm J, Vol. 16, no. 1, 2012, pp. 35-38
4. Eysenbach, G. **Consumer health informatics,** BMJ, Vol. 320, no. 7251, 2000, pp. 1713-1716

5. IRES (Institutul Român pentru Evaluare şi Strategie), **Românii şi internetul. Studiu privind utilizarea Internetului şi comportamentul internautic al românilor 2011,** accesed 15.05.2011, http://www.ires.com.ro/uploads/articole/ires_romanii_si_internetul_2011.pdf
6. McNally, S.L., Donohue, M.C., Newton, K.P., Ogletree, S.P., Conner, K.K., Ingegneri, S.E. and Kagnoff, MF. **Can Consumers Trust Web-Based Information about Celiac Disease? Accuracy, Comprehensiveness, Transparency, and Readability of Information on the Internet,** Interact J Med Res, Vol. 1, no. 1, 2012, pp. e1. doi: 10.2196/ijmr.2010
7. Nadaşan, V. **O evaluare a calităţii informaţiilor medicale din spaţiul virtual românesc,** PhD Thesis, University of Medicine and Pharmacy Targu Mures, Romania, 2011. p. 128
8. Nadaşan, V., Vancea, G., Georgescu, P.A., Tarcea, M. and Abram, Z. **The credibility, completeness and accuracy of information about first aid in case of choking on the Romanian Websites**, Journal of Applied Quantitative Methods, Vol. 6, no. 3, 2011, pp. 18–26
9. Seybert, H. **Internet use in households and by individuals in 2011,** Eurostat. European Commission, Statistics in focus, no. 66, 2011, accessed 04.05.2016 http://ec.europa.eu/eurostat/documents/3433488/5579964/KS-SF-11-066-EN.PDF/090e071f-c3a9-45d8-aa90-9b142251fd3a
10. Wang, L., Wang, J., Wang, M., Li, Y., Liang, Y. and Xu, D. **Using Internet search engines to obtain medical information: a comparative study,** J Med Internet Res, vol. 14, no. 3, 2012, pp. e74, doi: 10.2196/jmir.1943.
11. Weaver, III J.B., Thompson, N.J., Weaver, S.S. and Hopkins G.L. **Healthcare nonadherence decisions and internet health information,** Computers in Human Behavior, vol. 25, 2009, pp. 1373-1380, doi: 10.1016/j.chb.2009.05.011

[1] Corresonding author

WEIGHTING METHOD FOR DEVELOPING COMPOSITE INDICES. APPLICATION FOR MEASURING SECTORAL SPECIALIZATION

Ana Maria SAVA

Bucharest University of Economic Studies, Romania

E-mail: anamaria.sava89@yahoo.com

Abstract

When building a composite index, one might desire attributing different weights to factors whose influence it aggregates. Deciding on what weight to allocate to each factor may prove to be a difficult task if there is no possibility of finding an independent variable for the construct one tries to quantify. The present paper proposes a twist in using Principal Component Analysis as means for determining the weights of multiple factors based on which an index may be created. One example where finding an independent variable is not an option might be: developing an index for measuring sectoral specialization. Although over the years several instruments for measuring this construct have been developed, there is still no unanimous and universally accepted way of quantifying sectoral specialization and this paper designs a new index for measuring it by applying the weighting method advanced herein.

Keywords: *Weighting method; determining weights; composite index; PCA; no independent variable; sectoral specialization; Arts, entertainment and recreation*

1. Introduction

The goal of the present paper is to advance an instrument that enables building composite heterogeneity indices that takes into account the compound weighted influence of multiple factors considered to be relevant for assessing a certain phenomena. Following a previous approach in building such an index (Sava, 2016), this paper proposes a weighting method for designing indices in the absence of an independent variable. For exemplification reasons, the paper will present this method's functioning mechanism by developing a new sectoral specialization index (focusing on the local recreation industry). The paper will guide all the way from identifying the key factors whose influence will be taken into account in the index and determining the weight each factor should be attributed, to actually computing the index and discussing the output.

Over the years, economists have developed numerous instruments for measuring sectoral specialization. Palan (2010) divides them into two categories: specialization indices and heterogeneity indices. The first category – specialization indices – measure a country's absolute level of specialization, while the second – heterogeneity indices – measure the deviation of a country's industrial structure as compared to the average structure of a reference group of countries. Each of the two approaches presents both strengths and weaknesses.

Specialization indices give as output measures that can be interpreted per se, thus allowing focused application (measuring specialization only for the item of interest) and enabling straight-forward analysis of time series. However, their greatest disadvantage is that in their computation the development of other structures is not taken into consideration. On the contrary, heterogeneity indices use as benchmark exactly the average economic structure of the elements considered in the analysis. The major downside of these indices is that employing biased samples can generate wrongful results.

Acknowledging both advantages and disadvantages of the two categories of indices, and given the topic of interest for the present paper, the attention will be focused on the second category – heterogeneity indices. Examples of this type of indices are numerous: the Organization for Economic Co-operation and Development uses *Hannah-Kay Index* to highlight the sectoral composition measured on 20 industrial aggregates (OECD, 2013). European Central Bank quantifies sectoral specialization by means of *Krugman Index* that portrays the structure of a country's economy as compared to the EU structure (ECB, 2004). National Bank of Slovakia uses, next to Krugman index, the *Concentration Index* that shows for a given country a specific industry's contribution to the EU total and the *Lilien Indicator* that measures the speed of structural changes in employment (Čutková and Donoval, 2004).

Other heterogeneity indices mentioned by Palan (2010) and used in practice are: the *Index of Inequality in Productive Structure* that is similar to Krugman index but which grants large deviations an increased weight, *Relative Gini Index* extensively used both in concentration and specialization analyses, or the *Theil Index* that represents in fact a variation of the *Shannon Index* which establishes the employment level of a country in relation to that of the countries considered as reference group. Ioncică et al. (2010) proposed an index that allows calculating the degree of specialization for services sectors. The index proposed by them, called *Tertiary Specialization Index*, takes into account the share of services in GDP, employment and exports. As built, it can be applied to determine how specialized is the whole tertiary sector of a country, or one of its key service industries.

Among the aforementioned instruments, only the Tertiary Specialization Index is a composite index, but it does not imply assigning different weights to factors whose influence it aggregates. In this consisted also the main limitation of the prior variant of the specialization index that will be further presented (Sava, 2016), and for which purpose the weighting method that will be further presented was designed.

2. Description of the weighting method used

First and foremost, it is worth mentioning that the weights assigned through the method presented in this chapter are directly linked to the dataset used in the analysis, as is the computation of the heterogeneity index. Therefore, prior to explaining the weighting method, the coordinates of the index must be fixed.

Because the matter at hand implies assigning weights to factors for creating a composite index (therefore, not being able to rely on an independent variable to decide on factors' relevance for the measured construct), the accuracy of the results provided by the index depends primarily on the choice of factors. These have to be relevant for the studied concept and objectively chosen. The optimal choice of the indicators to be aggregated in the index implies that they can be measured on an ordinal scale, although it is not compulsory for this scale to have a fixed point of origin or an upper/lower limit.

Usual indicators for measuring sectoral specialization are: the sector's contribution to GDP, the share of employment of the sector in total employment, the size of the industry, the degree of concentration of companies in the sector, the share of exports, the government spending allocated to the sector, the value of private investments in the sector, or the expenditure for research and development activities in the sector. From this perspective, the index proposed herein will not deviate from the norm, as the following four factors were envisaged, each of them considered to have a positive influence on the degree of specialization or development of an economic sector:

- Gross production of the sector as share of GDP ($\frac{GP_S}{GDP}$);
- Employment in the sector as share of total employment ($\frac{E_S}{E}$);
- Sectoral government spending as share of total government spending ($\frac{G_S}{G}$);
- Household expenditure for products/services provided within the sector as share of the average shopping basket ($\frac{C_S}{C}$).

For exemplification reasons, the sector for which the sectoral specialization index proposed in this paper will be computed is one often neglected from similar studies – the Arts, entertainment and recreation sector.

All coordinates being established, data collection stage follows. For the current application, data was collected for 31 countries and for a timeframe of 11 years (starting from 2004 and ending with 2014, the year of the last available data). Data processing for running the analysis involves the calculation of normalized values for each of the four indicators used by applying Formula 1 for the individual samples of 31 countries, by treating each year separately.

$$X_t^n = \frac{X_t - \min(X_t)}{\max(X_t) - \min(X_t)} \tag{1}$$

where X_t^n is the normalized value of the factor ($\frac{GP_S}{GDP}$, $\frac{E_S}{E}$, $\frac{G_S}{G}$ or $\frac{C_S}{C}$);

X_t is the factor value at time *t*;

$\min(X_t)$ is the minimum value of the factor at time *t* within the sample;

$\max(X_t)$ is the maximum value of the factor at time *t* within the sample.

Only at this stage, the actual method of determining weights may be applied. The method consists of running a Principal Component Analysis and *restricting the number of components to be extracted to one*. In order to grasp which approach towards defining the weights is better, the PCA was run in two different manners:

- First approach: running the analysis on all data from all 11 years combined as to benefit from the robustness given by a large sample (SPSS output is presented in Table 1). In this case, the loads of each factor in the definition of the singular component represent the weights assigned to the factors.
- Second approach: grouping data by years and running the same analysis 11 times, once for each year and then averaging the results, by use of arithmetic mean, for each factor (SPSS output is displayed in Table 2). In this case, the weights assigned to the factors are represented by the average of the results obtained for the 11 analyses.

Table 1. First approach: PCA Output run on aggregated normalized data

Component Matrix[a]

	Component
	1
GP_s / GDP	,658
E_s / E	,800
G_s / G	,370
C_s / C	,753

Extraction Method: Principal Component Analysis.
a. 1 components extracted.
Source: Author's work

Table 2. Second approach: PCA Output run on breakdown-by-year normalized data

Component Matrix[a]

	Component											***Average***
	1											
Year	*2014*	*2013*	*2012*	*2011*	*2010*	*2009*	*2008*	*2007*	*2006*	*2005*	*2004*	
GP_s/GDP	,771	,727	,772	,773	,757	,700	,672	,475	,593	,592	,692	,684
E_s/E	,823	,836	,783	,754	,712	,747	,774	,867	,853	,883	,848	,807
G_s/G	,387	,388	,159	,141	,053	,085	,190	,583	,613	,539	,700	,349
C_s/C	,654	,628	,747	,768	,791	,795	,819	,824	,777	,797	,781	,762

Extraction Method: Principal Component Analysis.
a. 1 components extracted.
Source: Author's work

First observation is that all factors appear to be having a positive load in the definition of the component, therefore confirming the initial hypothesis – that each of them is considered to have a positive influence on the degree of specialization or development of the considered economic sector.

Moreover, both approaches towards defining the weights display very similar results, thus arguing for the robustness of the analysis. The biggest difference is recorded in the weight assigned to the third factor (sectoral government spending as share of total government spending) where the second method of calculation has displayed a result by 6% lower than the first method, while assigning the first factor (gross production of the sector as share of GDP) a load with 4% higher.

Table 3. Rescaling results as to obtain final weights

	Weight prior to rescaling	Weight after rescaling
GP_s / GDP	,684	,263
E_s / E	,807	,310
G_s / G	,349	,134
C_s / C	,762	,293
Sum	*2,602*	*1,000*

Source: Author's work

As the second method is considered more reliable (the input of elements in the sample is not multiplied artificially as each year is treated as a separate entity), the results obtained by the second approach will be used as weights in the calculation of the sectoral specialization index. Furthermore, in order to facilitate comprehension of the set of weights obtained, results were rescaled as to sum up to 1, and Table 3 shows the final set of weights used in computing the index.

3. Computation of the specialization index

Once the weights are set, the sectoral specialization index proposed herein can be calculated as a weighted average of the four factors (Formula 2). It ought to be noted that the specialization index is computed using the raw data, as opposed to the Principal Component Analysis that run on normalized data.

$$Sp = \frac{\sum x_i \times p_i}{\sum p_i} \qquad (2)$$

where Sp is the level of sectoral specialization;

x_i is the raw value (prior to normalization) of the factor ($\frac{GP_S}{GDP}$, $\frac{E_S}{E}$, $\frac{G_S}{G}$ or $\frac{C_S}{C}$);

p_i is the weight attributed to the factor ($\frac{GP_S}{GDP}$, $\frac{E_S}{E}$, $\frac{G_S}{G}$ or $\frac{C_S}{C}$).

In order for the results to be comparable, the last step in the computation of the index is to normalization the output, again by addressing each year separately. Therefore, the index can take values between 0 and 1, where proximity to 0 implies lack of specialization and proximity to 1 means a very high specialization degree.

Table 4. Sectoral specialization index computed for the Arts, entertainment and recreation sector (period 2004-2014)

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Austria	0,365	0,383	0,327	0,324	0,286	0,278	0,310	0,300	0,274	0,251	0,224
Belgium	0,151	0,173	0,124	0,080	0,062	0,028	0,086	0,083	0,085	0,049	0,049
Bulgaria	0,050	0,203	0,062	0,068	0,079	0,000	0,117	0,098	0,086	0,066	0,108
Croatia	0,605	0,553	0,518	0,488	0,319	0,390	0,393	0,296	0,320	0,319	0,262
Cyprus	0,474	0,541	0,479	0,496	0,413	0,408	0,476	0,477	0,472	0,431	0,367
Czech Republic	0,449	0,462	0,436	0,388	0,339	0,325	0,326	0,299	0,267	0,227	0,226
Denmark	0,518	0,530	0,470	0,430	0,373	0,381	0,379	0,369	0,349	0,325	0,316
Estonia	0,876	0,958	1,000	0,907	0,593	0,474	0,514	0,460	0,412	0,488	0,460
Finland	0,428	0,451	0,389	0,374	0,323	0,342	0,361	0,353	0,334	0,289	0,263
France	0,361	0,389	0,338	0,307	0,255	0,274	0,313	0,308	0,285	0,263	0,254
Germany	0,287	0,301	0,228	0,219	0,182	0,180	0,213	0,216	0,206	0,184	0,173
Greece	0,070	0,111	0,000	0,000	0,000	0,003	0,000	0,000	0,000	0,000	0,021
Hungary	0,412	0,442	0,367	0,324	0,282	0,249	0,333	0,312	0,293	0,272	0,270
Iceland	0,624	0,616	0,538	0,571	0,412	0,381	0,378	0,416	0,416	0,387	0,393
Ireland	0,309	0,334	0,341	0,322	0,309	0,286	0,302	0,311	0,296	0,253	0,222
Italy	0,164	0,157	0,073	0,057	0,063	0,068	0,139	0,107	0,106	0,079	0,069
Latvia	1,000	1,000	0,858	1,000	0,719	0,585	0,526	0,544	0,533	0,538	0,527
Lithuania	0,375	0,316	0,309	0,254	0,138	0,174	0,191	0,183	0,201	0,191	0,238
Luxembourg	0,112	0,228	0,069	0,089	0,069	0,035	0,067	0,071	0,045	0,037	0,041
Malta	0,400	0,520	0,540	0,866	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Netherlands	0,486	0,519	0,472	0,453	0,403	0,403	0,417	0,409	0,385	0,344	0,316
Norway	0,332	0,334	0,252	0,276	0,234	0,252	0,263	0,252	0,215	0,177	0,167
Poland	0,128	0,193	0,180	0,158	0,190	0,191	0,256	0,244	0,213	0,143	0,143
Portugal	0,164	0,202	0,125	0,068	0,061	0,027	0,081	0,067	0,053	0,009	0,000
Romania	0,000	0,000	0,007	0,004	0,041	0,052	0,162	0,164	0,149	0,120	0,178
Slovakia	0,257	0,310	0,246	0,222	0,158	0,210	0,262	0,287	0,282	0,254	0,268
Slovenia	0,487	0,528	0,494	0,446	0,365	0,330	0,403	0,331	0,311	0,238	0,259
Spain	0,519	0,554	0,530	0,507	0,429	0,386	0,419	0,393	0,317	0,288	0,301
Sweden	0,516	0,511	0,493	0,469	0,399	0,385	0,400	0,397	0,399	0,379	0,339
Switzerland	0,293	0,306	0,247	0,192	0,202	0,180	0,203	0,200	0,171	0,133	0,132
United Kingdom	0,463	0,499	0,462	0,433	0,374	0,337	0,339	0,330	0,326	0,283	0,281

Notes:

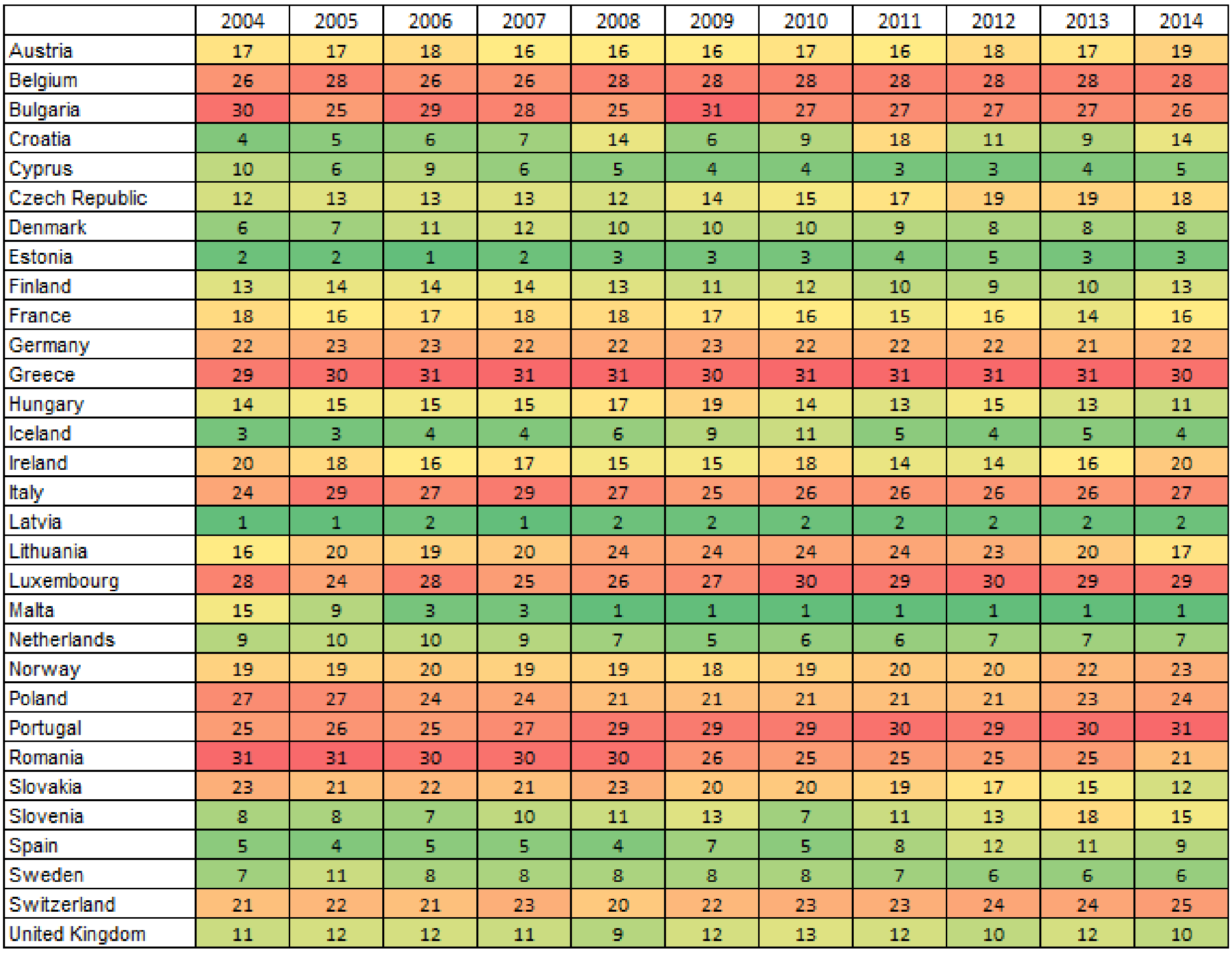
1) The horizontal bars are proportional with the values obtained for the index as compared to the entire sample of 31 countries taking into account the whole period of 11 years.

2) Background shades of green mark the differences in the evolution of the index for each country: a darker shade corresponds to higher values, while a lighter shade corresponds to lower values.

Source: Author's work, computed based on data retrieved from Eurostat (n.d.), Knoema (n.d.a), Knoema (n.d.b), INSEE (n.d.), INSSE (n.d.), NSI (n.d.), Statistics Iceland (n.d.), Statistics Norway (n.d.)

The analysis allows for observing a country's evolution across the years in terms of sectoral specialization. The trend can be regarded from two perspectives: evolution of scores or evolution in ranking. Tables 4 and 5 show the results obtained (scores and ranking) after computation of the sectoral specialization index, as defined herein. Both ways of looking at results present advantages in using them, but also require caution in interpreting the data.

Table 5. Countries' ranking considering the specialization index computed for the Arts, entertainment and recreation sector (period 2004-2014)



	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Austria	17	17	18	16	16	16	17	16	18	17	19
Belgium	26	28	26	26	28	28	28	28	28	28	28
Bulgaria	30	25	29	28	25	31	27	27	27	27	26
Croatia	4	5	6	7	14	6	9	18	11	9	14
Cyprus	10	6	9	6	5	4	4	3	3	4	5
Czech Republic	12	13	13	13	12	14	15	17	19	19	18
Denmark	6	7	11	12	10	10	10	9	8	8	8
Estonia	2	2	1	2	3	3	3	4	5	3	3
Finland	13	14	14	14	13	11	12	10	9	10	13
France	18	16	17	18	18	17	16	15	16	14	16
Germany	22	23	23	22	22	23	22	22	22	21	22
Greece	29	30	31	31	31	30	31	31	31	31	30
Hungary	14	15	15	15	17	19	14	13	15	13	11
Iceland	3	3	4	4	6	9	11	5	4	5	4
Ireland	20	18	16	17	15	15	18	14	14	16	20
Italy	24	29	27	29	27	25	26	26	26	26	27
Latvia	1	1	2	1	2	2	2	2	2	2	2
Lithuania	16	20	19	20	24	24	24	24	23	20	17
Luxembourg	28	24	28	25	26	27	30	29	30	29	29
Malta	15	9	3	3	1	1	1	1	1	1	1
Netherlands	9	10	10	9	7	5	6	6	7	7	7
Norway	19	19	20	19	19	18	19	20	20	22	23
Poland	27	27	24	24	21	21	21	21	21	23	24
Portugal	25	26	25	27	29	29	29	30	29	30	31
Romania	31	31	30	30	30	26	25	25	25	25	21
Slovakia	23	21	22	21	23	20	20	19	17	15	12
Slovenia	8	8	7	10	11	13	7	11	13	18	15
Spain	5	4	5	5	4	7	5	8	12	11	9
Sweden	7	11	8	8	8	8	8	7	6	6	6
Switzerland	21	22	21	23	20	22	23	23	24	24	25
United Kingdom	11	12	12	11	9	12	13	12	10	12	10

Note: Background colors highlight the position in the ranking, in relation to the entire sample of 31 countries taking into account the whole period of 11 years. Green shades correspond to upper positions of the ranking, while red shades mark lower positions.
Source: Author's work, computed based on data from Table 4

Each year, a country's scores are calculated relative to the other countries' individual performances; therefore a positive evolution of scores does not necessarily imply an increase in specialization (it might be just due to a decrease in other countries' performances). But, in the case where the competitive context remains broadly unchanged, such an approach offers a more contoured overview of the evolution. By focusing the attention on the ranking evolution, one might be tricked into thinking a country registers a striking increase in specialization, but if it is the case of outrunning countries with very close specialization levels, then an increase of less than 1% may generate such an outcome. However, outranking a long established country that consistently displayed a high degree of specialization may constitute a notable performance and looking at the ranking can become a good indicative.

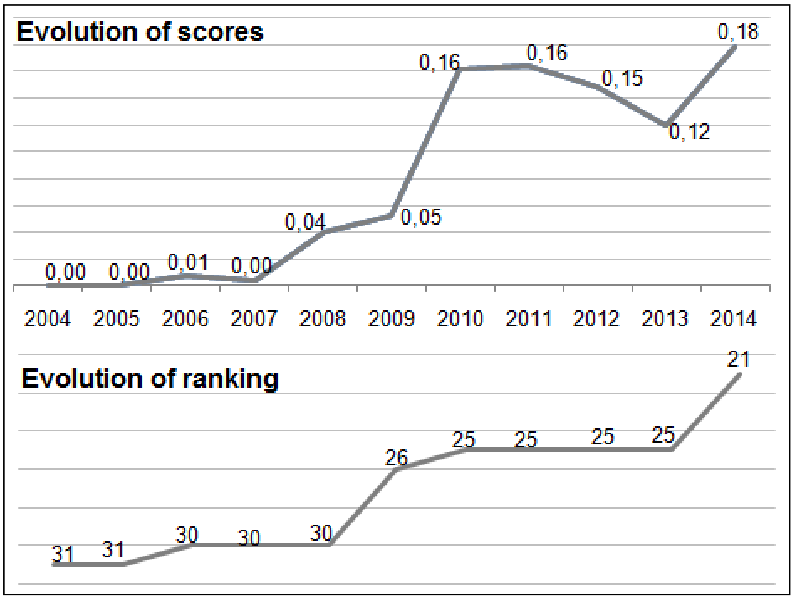


Figure 1. Romanian Arts, entertainment and recreation sector's specialization evolution (period 2004-2014)

Source: Author's work

For illustration purposes, a zoom-in on the results is depicted in Figure 1, where is presented Romania's evolution regarding the specialization of the Arts, entertainment and recreation sector. Both indicators' evolution converges to the same overall conclusion – that over time Romania's position has strengthened in European context by means of an increase in relative specialization. However, if one were to look only at the evolution of scores, the 2011-2013 scores might be misleading, as although the score has dropped, Romania maintained its position in the ranking due to the concomitant decrease in other countries' performance.

4. Concluding remarks

Although initially designed as a method for reducing large sets of factors into a more manageable number of components, the Principal Component Analysis is used herein as means of reducing a rather small number of factors into just one single component with the aim of using the component loads as weights for developing a composite index.

By following the steps described, this method may be applied to developing other indices (it is not bound to working only for specialization measurements). Furthermore, although in the present paper the specialization index was used for quantifying local recreation specialization, its application may be generalized to other economic sectors.

The main limitation of the study lies in the scarcity of data collected for the analysis. Although data was gathered generally from a single source (Eurostat), some additional figures were retrieved from other various online sources (data provided by national statistics institutes, online databases) and further data processing techniques as to obtain homogeneity were then applied. Even though the time period considered in the analysis is quite extended, there were several gaps in the data that had to be filled in by estimations. Moreover, sample size is rather small and all elements are concentrated in just one geographic region.

Because it is a heterogeneity index (implying that a country's results are obtained as a result of the structure of the reference group of countries), interpreting output ought to be carefully considered because both score evolution and ranking evolution can be misleading and may cause drawing biased conclusions. Therefore, competitive context should always be a concern in interpreting results.

References

1. Čutková, Z. and Donoval, M. **Sectoral specialization in the SR,** BIATEC, Vol. 10, No. 12, 2004, 5–7
2. European Central Bank, **Sectoral specialization in the EU,** A macroeconomic perspective, Occasional Paper Series, Vol. 19, 2004
3. Eurostat (n.d.) **Eurostat Database** [online], Accessed April 10, 2016, from http://ec.europa.eu/eurostat/data/database
4. Ioncica, M., Draghici, M., Petrescu, C. and Ioncica, D. **Services specialization (a possible index) and its connection with competitiveness: the case of Romania,** The Service Industries Journal, Vol. 30, No. 12, 2010, pp. 2023-2044
5. Knoema (n.d.a) **Croatia - Recreation and culture: share of household expenditure (COICOP 09) (%)** [online], Accessed April 10, 2016, from https://knoema.com /cpc_ecnacoi/candidate-countries-and-potential-candidates-annual-national-accounts-breakdown-of-final-consumption?tsId=1000090
6. Knoema (n.d.b) **Final consumption expenditure of households** [online], Accessed April 10, 2016, from https://knoema.com/SNA_TABLE5_2015/final-consumption-expenditure-of-households?country=1000300-switzerland
7. National Institute of Statistics and Economic Studies of France – INSEE (n.d.) **Quarterly payroll employment by sector - Main groups** [online], Accessed April 10, 2016, from http://www.bdm.insee.fr/bdm2/choixCriteres?codeGroupe=1185
8. National Institute of Statistics of Romania – INSSE (n.d.) **Tempo Online** [online], Accessed April 10, 2016, from http://statistici.insse.ro/shop/?lang=en
9. OECD **OECD science, technology and industry scoreboard 2013. Innovation for growth,** 2013 [pdf]. Accessed April 10, 2016, from http://www.oecd. org/sti/scoreboard-2013.pdf
10. Palan, N. **Measurement of Specialization – The Choice of Indices,** FIW - Working Paper, No. 62, December 2010
11. Republic of Bulgaria National Statistical Institute – NSI (n.d.) **Sector General Government** [online], Accessed April 10, 2016, from http://www.nsi.bg/en/content/ 5267/sector-general-government
12. Sava, A.M. **Correlating Local Recreation Specialization to Prosperity: Study on European Union Countries,** Entrepreneurship, Business and Economics, Vol. 2, 2016, pp. 319-330
13. Statistics Iceland (n.d.) **Economy** [online], Accessed April 10, 2016, from http://px. hagstofa.is/pxen/pxweb/en/Efnahagur/Efnahagur__fjaropinber__fjarmal_opinber__fjarmal_opinber/?rxid=a5ce19c0-4600-4c6c-9dce-8f726e07869a
14. Statistics Norway (n.d.) **Survey of consumer expenditure** [online], Accessed April 10, 2016, from https://www.ssb.no/statistikkbanken/SelectVarVal/Define.asp? MainTable=UtgHusType&KortNavnWeb=fbu&PLanguage=1&checked=true