Volume 5 Issue 1

ISSN: 2321-1857

International Journal of

Business Analytics & Intelligence

April 2017

A Publication of Publishing India Group

International Journal of Business Analytics & Intelligence

Editorial Message

Happy New Year!!!

It is a great pleasure for me to make available to you the 1st issue of Volume 5 in this year 2017.

We are continuously putting our effort in developing the alliance between academia, industry practitioners to bring various perspective of data science so as to enhance readership. The journal is a platform for exchanging research insights, analytical techniques and knowledge in various areas which include but are not limited to a destination, yet we emphasis on constant expedition.

In this issue of IJBAI, we are pleased to publish six insightful and informative papers which are focused on the theme of multiple techniques on data processing, sampling, measurement and data mining techniques. Environmental concerns have been on the agendas of industry and academia for more than decades and we are happy to bring a perspective on the theme of Green Advertisement. You will receive an informative research framework which signifying antecedents to skepticism toward green advertising and description of an interpretive structural modeling technique to conceptualize the inter-relationship between/among variables. The adverse effect of climate change is one alarming area in today's world and by addressing the challenge; you will read the paper on 'Structural Equation Modeling to determine the climate change' brings the in depth understanding of the need to inclusion of the topic in management course.

In the era of data science and machine learning, we brought an insightful subject matter by describing circular data analysis distribution of Traffic Accident Times in India. On computation techniques, one interesting paper focuses on a faster, flexible and scalable Text Analytics Framework using Apache Spark and Combination of Lexical and Machine Learning Techniques to assess people's mindset, opinion and feedback. This paper is industry agnostic and useful to the extent of sentiment prediction with cost effective techniques, hence ready to adapt for various institutions. A comparative analysis of different measurement scales and an effective framework to perform analytics over the voluminous dataset, with Progressive Sampling Model to reduce the time yet maintaining the quality of the result at high level are the two key papers on measurement techniques.

I would like to thank the researchers and renowned data science practitioners who have honored us by choosing our young journal to publish some of their research. I would like to thank the collaboration of several referees, whose excellent and anonymous work has made possible the publication of this issue.

I am sure that our readers will enjoy and learn a lot from the present issue. Do let us know your wish, suggestions and views to enrich our journal

Sincerely yours,

Madhumita Ghosh
Joint Editor-in-Chief
Dated: 27th January, 2017

International Journal of Business Analytics and Intelligence

Volume 5 Issue 1 April 2017

ISSN: 2321-1857

Perspectives

Research Papers

Journal is also available online at www.publishingindia.com

Article can be accessed online at http://www.publishingindia.com

Application of Derivatives to Nonlinear Programming for Prescriptive Analytics

Tuhin Chattopadhyay, Ph. D.

Gone are the days when business analytics would bank on statistics alone. Besides the traditional probability theory and statistics, the machine learning techniques of the present era, work in complete sync with linear algebra, graph theory, dynamic programming, multivariate calculus etc. As far as multivariate calculus is concerned, the different methods that lend support to machine learning algorithms are differential and integral calculus, partial derivatives, gradient and directional derivative, vector-valued function, Jacobian matrix and determinant, Hessian matrix, Laplacian and Lagrangian distributions etc. The present article will discuss the applications of second order derivatives and partial derivatives on optimization problems, as required for prescriptive analytics.

Prescriptive analytics provides precise decisions on the course of action that the business will undertake for success in future. One of the prominent applications of prescriptive analytics in marketing is the optimization problem of marketing budget allocation. The business problem is to figure out the optimum quantity of budget that needs to be allocated from the total advertising budget to each of the advertising media like TV, press, internet video etc. for maximizing the revenue. The budget optimization problem is solved either through Linear or Nonlinear Programing (NLP) which depends on whether

i. The objective function is linear/ nonlinear
ii. The feasible region is determined by linear/nonlinear constraints.

Thus, one of the important assumptions for linear programming is the constant returns to scale for each of the advertising media. But the real world TV advertisement data, as plotted in Figure 1, defies such assumption as the graph shows a concave function. The constraint that may be considered for developing such optimization problem is the maximum amount that should be spent on a particular media such that beyond that point any further expenditure may lead to the increase in revenue but at a decreasing rate. Thus, it's important to find out the diminishing point of return for each of the advertising medium. Figure 1 provided below shows the revenue generated against the cost incurred for TV advertisement. Both the cost and the revenue mentioned in the current paper are dollar value in thousands.

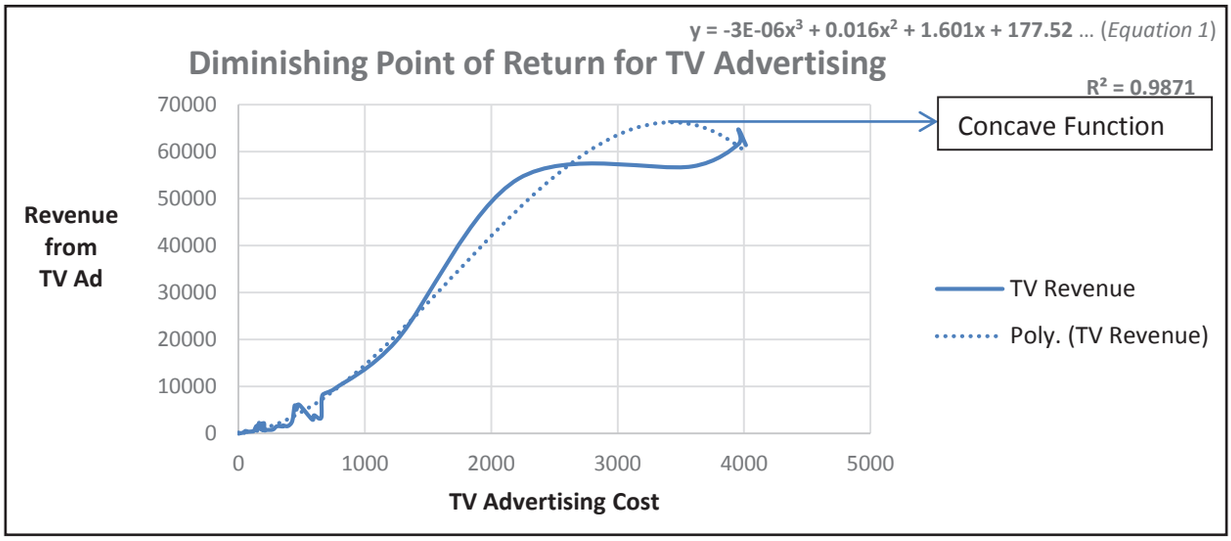


Fig. 1: Diminishing Point of Return for TV Advertisement

The curve that best fits the plotted revenue and cost of TV advertisement is cubic and is plotted in *Figure 1* and mentioned in *Equation 1*. The R^2 achieved through the cubic equation is a whopping 98.7%. The first and second order derivatives of the cubic equation are calculated as follows:

Polynomial Equation, $y = -3E\text{-}06x^3 + 0.016x^2 + 1.601x + 177.52$ … (*Equation 1*)

First Derivative of Equation 1, $\frac{dy}{dx} = -9.00E\text{-}06x^2 + 0.032x + 1.601$… (*Equation 2*)

Second Order Derivative of Equation 1, $\frac{d^2y}{dx^2} = -1.80E\text{-}05x + 0.032$… (*Equation 3*)

The inflection point is identified where the second derivative changes from positive to negative. Therefore, mathematically, it's the point where second derivative is *0*. Therefore, solving *Equation 3,* the cost at diminishing point of return is *1777.78* and the corresponding revenue at the diminishing point of return after plugging the value of *x* in *Equation 1* is *36735.68*. For further studies on the application of second derivatives on non-linear optimization, the readers may refer to Newton–Raphson algorithm and conjugate direction algorithms.

Partial derivative is the other prominent application of calculus on optimization problems Partial derivatives of a function with several variables are accomplished when a particular variable's derivative is computed keeping other variables constant. One of the most widely used applications of partial derivative is the least square criterion where the objective is to find out the best fitting line by minimizing the distance of the line from the data points. This is achieved by setting first order partial derivatives of the intercepts and the slopes equal to zero. The other major applications of partial derivatives are as follows:

i. The second order partial derivative is used in optimization problem to figure out whether a given critical point is a relative maximum, a relative minimum, or a saddle point.
ii. The assignment of penalty for conversion of an optimization problem to an unconstrained optimization problem through sequential unconstrained minimization technique.

The approaches discussed above show how calculus can be integrated with nonlinear programming while delivering an optimization solution. In the same manner, Lagrangean based techniques can also be integrated with Mixed Integer Non-Linear Programming (MINLP) to provide with the marketing budget optimization solution. Thus, in the present era, the data scientists cannot bank on a single technique to provide analytics solution. The real challenge is to figure out how multiple techniques can be creatively amalgamated to provide a solution as unique as the business problem.

Article can be accessed online at http://www.publishingindia.com

Exploring Skepticism Toward Green Advertising: An ISM Approach

Vibhava Srivastava*

Abstract

There has been almost the analogous evolution of the phenomenon of green marketing and advertising across the globe though same is not so evident in consumption of environment friendly/green products. The premise of this research is centered upon the domain of green advertising wherein individuals though environment-conscious, tend to be skeptical about green advertising subsequently affecting their attitude and furthering their intention to purchase and consume green products. A group of ten experts with varied background were pooled in and introduced to the variables, identified through an extensive literature review, for exploring the contextual relationships. Interpretive Structural Modeling (ISM) and MICMAC analysis was further employed to conceptualize the inter-relationship between/among these variables. Based on driving power and dependence of respective variable, a framework signifying antecedents to skepticism toward green advertising, is conceptualized and proposed. The present study consolidates and suggests a conceptual framework to understand antecedents leading to skepticism towards green advertising. It is an exploratory research and not the conclusive. Though it proposes a research framework but the same is subject to empirical investigation.

Keyword: Green Advertising, Skepticism, Environment consciousness, ISM

INTRODUCTION

The concept of green marketing has evolved overly since the time it was first introduced in the late 1980s (Peattie & Crane, 2005) and so there has been the analogous growth in green advertising. Consumers' concern about the environment has increased considerably in recent years and the development of green products and the concomitant use of green advertising have continued to grow (Royne et al., 2012). With a higher consumer awareness of environmental issues, many companies across globe, have jumped on the bandwagon by adopting overtly "green" strategies (Ginsberg and Bloom, 2004; Laroche et al., 2001; Polonsky and Rosenberger, 2001), often making environmental claims in their advertising campaigns with the aim of gaining an edge over their competitors (Connolly and Prothero, 2003; Banerjee et al., 1995; Carlson et al., 1993).

However various researches show that the evolved environmental consumerism is not keeping pace with the increasing number of consumers who report that they are very concerned about the environment (Roper Consulting, 2010) which implies that one's environment consciousness/concern is not resulting into one's consumption of such products i.e. green. There has been a serious dilemma for marketers desiring to target the environmentally conscious/green consumer who is somewhat cynical about marketing activities and is likely to discount particularly, advertising messages and to distrust corporate motives (Zinkhan & Carlson, 1995). Shrum et al., (1995), report that green consumers, though are most likely to buy green products but also are more skeptical of advertising in general and are not brand loyal though past research has identified environmental consciousness as an influential factor not only affecting consumer responses to green advertisements (Schuhwerk & Lefkoff-Hagius 1995) but also determining advertising theme selection (D'Souza & Taghian 2005). Carlson et al., (1993) too report that many green advertising claims are both vague and ambitious. Marketers have a tendency to "push the envelope," especially when promoting

* Assistant Professor (Marketing), Management Development Institute, Gurgaon, Haryana, India.
Email: vibhava.srivastava@mdi.ac.in

products in new media or when creating new types of appeals e.g., green appeals (Zinkhan & Carlson, 1995). Consumers may be suspicious about green advertisers and their overly ambitious claims. It has also been noted that consumers often feel confused about the environmental claims in advertisements (Mayer et al., 1992).

Despite the popularity of the environmental movement and the recent increase in consumer environmental concern, green products, and green advertising, the actual purchasing of green products and patronage of green firms is not representative of the level of self-professed environmental concern by consumers (Gregory & Di Leo, 2003; Mayer et al., 1992; Shrum et al., 1995). Advertising claims that are difficult for consumers to verify are likely to prompt skepticism, consumer distrust, or disbelief of marketer actions (Foreh & Grier, 2003). Not surprisingly, environmental claims are often viewed skeptically and are miscomprehended (Beltramini & Stafford, 1993; Carlson et al., 1993; Shrum et al., 1995). Consumers with very high levels of environmental skepticism would be difficult to persuade with advertisements emphasizing the environmental benefits of an eco-friendly product (Royne et al., 2012). Interestingly, a wealth of research has examined potential factors that might affect the effectiveness of green advertising campaigns (e.g. Obermiller, 1995; Schuhwerk & Lefkof-Hagius, 1995; Zinkhan & Carlson, 1995; Chan, 2000; Hartmann & Apaolaza-Ibáñez, 2009).

The present study is exploratory in nature wherein author has tried to conceptualize and propose a framework signifying antecedents to skepticism toward green advertising, using Interpretive Structural Modeling (ISM). An extensive literature review was carried out to identify various antecedents of skepticism toward green advertising. ISM was employed to conceptualize the inter-relationship between/among antecedents and the skepticism toward green advertising. Subsequently MICMAC analysis of developed ISM model was carried out and various propositions weremade. Finally the research and managerial implications were discussed.

Theoretical Underpinning

Extensive review of literature using key words like green marketing, advertising and skepticism etc.,was carried out and as a result, number of variables were identified, which are shown in table 1.

Skepticism toward Green Advertising

Skepticism toward advertising in general was addressed by different researchers in past, where in efforts were also made to define the same (Kanter & Wortzel, 1985; Lutz, 1985; MacKenzie & Lutz, 1989; Ford et al., 1990; Boush et al., 1994; Friestad & Wright, 1995; Calfee & Ringold, 1994; Mangleburg & Bristol, 1998; Obermiller & Spangenberg, 1998, 2005). The consumer's lack of trust in advertising has been identified as a common thread across various definitions of Skepticism toward advertising (Boush et al., 1994; Mangleburg & Bristol, 1998; Obermiller & Spangenberg, 1998), along with dimensions like mistrust of advertiser motives and disbelief in advertisement (Boush et al., 1994), the perceived motivation of advertisers as well as the claims made by them (Mangleburg & Bristol, 1998) thus suggesting Ad Skepticism as a multi-dimensional construct. However Obermiller and Spangenberg (1998) found it as uni-dimensional construct referring to consumers' tendency toward the disbelief of advertising claims. A clear distinction has also been made between Ad Skepticism and the general attitude toward advertising (Obermiller & Spangenberg, 1998). Similarly, the message source which is contextual and manipulable, should not be considered a dimension of Ad Skepticism (Hardesty et al., 2002). Thus, Advertising claims that are difficult for consumers to verify are likely to prompt skepticism, consumer distrust, or disbelief of marketer's actions (Foreh & Grier, 2003).In continuation to the same, environmental claims are often viewed skeptically and are miscomprehended (Beltramini & Stafford, 1993; Carlson et al., 1993; Shrum et al., 1995; Bickart & Ruth, 2012). Thus, skepticism toward green advertising can be referred as consumer's tendency toward the disbelief of claims in green advertising and can have dimensions like disbelief of environmental/ green claims as well as mistrust of advertiser' motives.

Environment Consciousness

Environmental consciousness is defined as a psychological tendency to engage in pro-environmental behavior that reflects the individual's recognition of, and value judgments and behavior intentions towards, environmental issues (Schlegelmilch et al., 1994; Zelezny & Schultz, 2000; Zheng, 2010). Invariably it has been found to be driving varied purchasing and consumption behavior of individuals (Roberts, 1996)

including the likely assessment of post-consumption impact on environment by individuals (Zinkhan & Carlson, 1995) thus limiting the available consumption choices for individuals (Sherif et al., 1965). Thus the degree of environment consciousness has some impact on green consumerism. Green consumers were found to be skeptical about advertisers' environmental/green claims as per the various researches conducted in developed countries particularly (Shrum et al. 1995; Zinkhan & Carlson 1995; Manrai et al. 1997; Mohr et al. 1998) thus suggesting the likely relationship between the degree of environment consciousness and skepticism toward green advertising.

Green Issue Proximity

Green issue proximity refers to the degree of closeness between an individual and the green/environmental issue(s) which is evaluated on the basis of spatial distance (Chang, 2012). The proximity or low and high degree of closeness with green/environmental issue(s) has likely impact upon individuals' sensitivity towards such problems (Takács-Sánta, 2007) thus having varied perceptions in this context (Kollmuss & Agyeman, 2002) particularly when the change in environment is incremental and not immediately tangible (Kollmuss & Agyeman, 2002). The social impact theory (Latané 1981; Latané & Bourgeois 2001) indicates the possible impact of issue proximity, in forming the perceptions and subsequently, in assessing the effectiveness of green marketing/advertising (Chang, 2012).

Credibility of Green Messages

The credibility of any advertisement in general refers to the positive and favorable perception formed by individual(s)/audience(s)/consumer(s) about the claims made to be truthful, convincing and believable in the said advertisement (MacKenzie & Lutz, 1989; Kim & Damhorst, 1999; Goldsmith et al., 2000). The message is perceived to be credible if the same is trustworthy (Szymecko, 2003). The credibility of green messages is made to establish through different means like eco-labeling (Nimon & Beghin, 1999; Thorgersen, 2002), elaborating product specification particularly environmental attributes (Buda & Zhang, 2000), emphasizing recycling credentials (Karna et al., 2001) and establishing communicators' credibility (Szymecko, 2003) etc. However, in general the credibility of green advertising is considered to be relatively low (do Paco & Reis, 2012) thus it might lead to skepticism toward green advertising.

Perceived Brand Credibility

Individuals' perceived brand credibility is the believability of the product position information contained in a brand, which entails consistently delivering what is promised (Erdem & Swait, 2004). It leads to various outcomes like positive and favorable consumers' attitudes (Goldsmith et al., 2000), increase in customer loyalty (Ginsberg & Bloom, 2004) and increase in sales (Marshall & Mayer, 1992) etc. It is pertinent to say that brand(s) is not only an important and rich source of information to consumer(s) for making purchase (Branthwaite, 2002) and consumption decision but also a source of competitive advantage for firm(s) (Grace and O'Cass, 2002). In this regard Goldsmith et al., (2000) postulate that higher levels of perceived deception were found to be associated with lower levels of perceived brand credibility thus forming less favorable attitude toward the advertisement as well as the advertised brand.

Attitude toward Green Advertiser

Individuals' attitude toward the advertiser in general, is defined as a learned predisposition to respond in a consistently favorable or unfavorable manner toward the sponsoring organization (MacKenzie & Lutz, 1989) and is supposedly contributing in the formation of one's attitude toward the advertisement (MacKenzie & Lutz, 1989). Advertisers have relied on both informational and emotional appeals to help form and change attitudes and to convince consumers to purchase (Edell & Burke, 1987; Ratchford, 1987; Rossiter et al., 1991). The resultant perceived credibility of an advertiser determines the extent to which the audience perceives the claims made to be truthful and believable (Kim & Damhorst, 1999; Goldsmith et al., 2000; Phau & Ong, 2007). Thus individuals' attitude toward green advertiser can be interpreted as individual's perception formed on the basis of truthfulness and believable green claims made in the said advertisement.

Green Ad Guilt Appeal

Appeal(s) as tool in advertising, with the elements of guilt, viewed as complex negative emotional reaction (Festinger, 1962; Ghingold, 1981) on the part of individual(s), reportedly influences individuals' attention, attitude(s) toward product(s), and intention(s) to purchase (Coulter & Pinto 1995; Huhmann & Brotherton 1997; LaBarge & Godek 2006; Basil et al. 2006, 2008; Hibbert et al. 2007; Hill & Moran 2011; Chang 2011). However, there has been inconsistency and contradiction reported in various research with respect to the guilt appeal(s) whether or not same works or backfires (e.g. Coulter & Pinto 1995; Cotte et al., 2005; Turner et al. 2009). In the context of green advertising, same has been reportedly used to influence consumer behavior (Banerjee et al., 1995; Huhmann & Brotherton 1997; Hibbert et al., 2007; Chang, 2011) thus it can be referred as one of the determinants of attitude towards green advertising, skepticism particularly.

Table 1: Variables Identified

Sr. No.	*Variables*	*Author(s)*	*Inferences*
1.	Skepticism toward Green Advertising	Carlson et al. (1993); Beltramini & Stafford (1993); Boush et al. (1993), 1994; Shrum et al. (1995); Mangleburg & Bristol (1997); Mohr et al. (1997); Obermiller & Spangenberg (1997);Crane (2000); Laroche et al. (2001); Karna et al. (2001); Forehand & Grier (2003); Cone 2011; Finisterra do Paco & Reis (2012); Royne et al. (2012)	Tendency toward disbelief of ad claims. It often refers to the consumer's lack of trust in advertising. Consumers with very high levels of environmental skepticism would be difficult to persuade.
2.	Environment Consciousness	Zinkhan & Carlson (1995); Shrum et al. (1995); Schlegelmilch et al. (199); Roberts (199); Manrai et al. (1996); Mohr et al. (1997); Zelezny & Schultz (2000); Cotte et al. (2005); Zheng (2010)	A psychological tendency to engage in pro-environmental behaviors that reflects the individual's recognition of, and value judgments and behavior intentions towards, environmental issues.
3.	Green Issue Proximity	Latané (1971); Latané & Bourgeois (2001); Kollmuss & Agyeman (2002); Takács-Sánta (2006); Chang (2012)	The degree of closeness between a person and the issue. People tend to be less sensitive to environmental problems arising far from their own place of living.
4.	Credibility of Green Messages	MacKenzie & Lutz (1979); Nimon & Beghin (1999); Kim & Damhorst (1999); Goldsmith et al. (2000); Buda & Zhang (2000); Karna et al. (2001); Thorgersen (2002); Szymecko (2003); Phau & Ong (2006)	The extent to which the consumer perceives claims made about the brand in the ad to be truthful and believable. Important factors that will influence the response to an advertisement are the manner in which the message is framed and the perceived credibility of the source.
5.	Brand Credibility	Goldsmith et al. (2000); Grace & O'Cass (2002); Branthwaite (2002); Ginsberg & Bloom (2004)	The higher levels of perceived deception were associated with lower levels of perceived credibility, and with less favorable attitudes toward the advertisement and the advertised brand.
.6.	Attitude toward Green Advertiser	MacKenzie & Lutz (1979)	A learned predisposition to respond in a consistently favorable or unfavorable manner toward the sponsoring organization.
7.	Green Ad Guilt Appeal	Ghingold (1971); Banerjee et al. (1995); Coulter & Pinto (1995); Huhmann & Brotherton (1996); Cotte et al. (2005); Block (2005); LaBarge & Godek (200); Basil et al. (200); Hibbert et al. (2006); Basil et al. (2007); Turner et al. (2009); Hill & Moran (2011); Chang (2011), (2012)	A complex emotional reaction on the part of individuals and can be seen as a combination of negative emotions, such as regret, remorse and self-blame. Guilt appeals may work or backfire.

Conceptualizing Framework through Interpretive Structural Modeling (ISM)

In order to conceptualize the framework signifying antecedents of Skepticism toward Green Advertising, Interpretive Structural Modeling (ISM) was utilized. ISM, first proposed by J. Warfield (1974, 1976), is an interactive learning process in which a set of different and directly related variables affecting the issue under consideration are structured into a comprehensive systemic framework. It can act as a tool for imposing order and direction on the complexity of relationships amongst elements pertaining to an issue (Sage, 1977; Watson, 1978). It has been used in varied contexts, to explore a wide range of issues (e.g. Saxena & Vrat, 1992; Mandal & Deshmukh, 1994; Singh et al., 2003; Ravi & Shankar, 2004; Bolanos et al., 2005; Huang et al., 2005; Thakkar et al., 2006; Agarwal et al., 2007; Srivastava & Singh, 2010; Pfohl et al., 2011 etc.). In ISM, no knowledge of the underlying process is required for the participants rather a basic understanding of the subject is sufficient enabling them to respond to the series of relational queries (Srivastava & Singh, 2010). ISM guides and records the results of group deliberations on complex issues in an efficient and systematic manner thus producing a structured model or graphical representation of the original problem situation that can be communicated more effectively to others (Sage, 1977; Watson, 1978). The subsequent section deals with the ISM process where in the various sequential steps of the ISM process, are explicated.

ISM starts with formation of the group of people with relevant knowledge, skills and backgrounds. In this study for identifying the contextual relationship among the variables ten experts were pooled in and consulted who are well conversant with the green marketing practices, particularly green advertising. The group had equal representation from industry as well as academia.These experts were, then introduced to the variables which had beenIdentified and selected for the study through an extensive literature review. The variables are listed in Table 2.

The expert group was asked to deliberate upon and suggest the pair-wise relationships amongst the variables in order to develop the Structural self-interaction Matrix (SSIM). On the basis of possible contextual relationship for each variable, the existence of a relationship between any two variables and the associated direction of the relationship was questioned. Four symbols were used to denote the direction of relationship between the variables (a & b):

Table 2: List of Variables

Variables	
V1	Environment Consciousness
V2	Green Issue Proximity
V3	Credibility of Green Messages
V4	Skepticism toward Green Advertising
V5	Brand Credibility
V6	Attitude toward Green Advertiser
V7	Green Ad Guilt Appeal

V: a is related to b but b is not related to a.

A: a is not related to b but b is related to a.

X: a and b both are related to each other.

O: a and b both are not related to each other.

The Structural self-interaction Matrix (SSIM), thus developed is shown in Table 3.

Table 3: Structural Self-interaction Matrix (SSIM)

Variables	*7*	*6*	*5*	*4*	*3*	*2*
V1 Environment Consciousness	O	V	O	V	V	A
V2 Green Issue Proximity	O	V	O	V	V	
V3 Credibility of Green Messages	A	A	X	V		
V4 Skepticism toward Green Advertising	A	A	A			
V5 Brand Credibility	O	A				
V6 Attitude toward Green Advertiser	A					
V7 Green Ad Guilt Appeal						

The SSIM was transformed into a binary matrix, known as reachability matrix by substituting V, A, X, and O by 1 and 0 as per the case. The rules for the substitution of 1's and 0's are as follows:

V If a is related to b but b is not related to a, then a→b=1, b→a=0.

A a is not related to b but b is related to a, then a→b=0, b→a=1.

X a and b both are related to each other, then a→b=1, b→a=1.

O a and b both are not related to each other, then a→b=0, b→a=0.

Reachability matrix thus developed, was checked for transitivity. It states that if a variable A is related to B and B is related to C, then A is necessarily related to C (Table 4). In this table, the driving power and dependence of each variable was also shown. The driving power of a particular variable is the total number of variables (including the said variable itself) which it may help to achieve. The dependence is the total number of variables which may help in achieving it.

Table 4: Reachability Matrix

Variables	1	2	3	4	5	6	7	Driver Power
V1	1	0	1	1	0	1	0	4
V2	1	1	1	1	0	1	0	5
V3	0	0	1	1	1	0	0	3
V4	0	0	0	1	0	0	0	1
V5	0	0	1	1	1	0	0	3
V6	0	0	0	1	1	1	0	3
V7	0	0	1	1	0	1	1	4
Dependence	2	1	5	7	3	4	1	

The reachability matrix was decomposed into different levels in order to create structural framework. The reachability and antecedent set for each variable was found out from the reachability matrix where in the reachability set for a particular variable consists of the variable itself and the other variables, which it may help to achieve while the antecedent set consists of the variable itself and the other variables, which may help in achieving it.

Subsequently, the intersection of these sets was derived for all variables. The variable for which the reachability and the intersection sets are the same is assigned the top-level in the ISM hierarchy, which would not help achieve any other variable above its own level. After the identification of the top-level variable, it was discarded from the other remaining variables. From Table 5, it can be seen that skepticism toward green advertising (V4) was found at Level I. Thus, it would be positioned at the top of the ISM framework. This iteration was continued till the levels of each variable were found out. The identified levels aid in building the digraph and the final framework of ISM. The variables, along with their reachability set, antecedent set, intersection set and the levels, are shown in Tables 5-9.

Table 5: Iteration 1

Variables	*Reachability Set*	*Antecedents Set*	*Intersection Set*	*Levels*
1	1,3,4,6	1,2	1	
2	1,2,3,4,6	2	2	
3	3,4,5	1,2,3,5,7	3,5	
4	4	1,2,3,4,5,6,7	4	I
5	3,4,5	3,5,6	3,5	
6	4,5,6	1,2,6	6	
7	3,4,6,7	7	7	

Table 6: Iteration 2

Variables	*Reachability Set*	*Antecedents Set*	*Intersection Set*	*Levels*
1	1,3,6	1,2	1	
2	1,2,3,6	2	2	
3	3,5,	1,2,3,5,7	3,5	II
5	3,5	3,5,6	3,5	II
6	5,,6	1,2,6	6	
7	3,6,7	7	7	

Table 7: Iteration 3

Variables	*Reachability Set*	*Antecedents Set*	*Intersection Set*	*Levels*
1	1,6	1,2	1	
2	1,2,6	2	2	
6	6	1,2,6	6	III
7	6,7	7	7	

Table 8: Iteration 4

Variables	*Reachability Set*	*Antecedents Set*	*Intersection Set*	*Levels*
1	1	1,2	1	IV
2	1,2	2	2	
7	7	7	7	IV

Table 9: Iteration 5

Variables	*Reachability Set*	*Antecedents Set*	*Intersection Set*	*Levels*
2	2	2	2	V

This provided a multi-level interpretive structural framework in which the relations amongst variables were clarified (Figure 1).

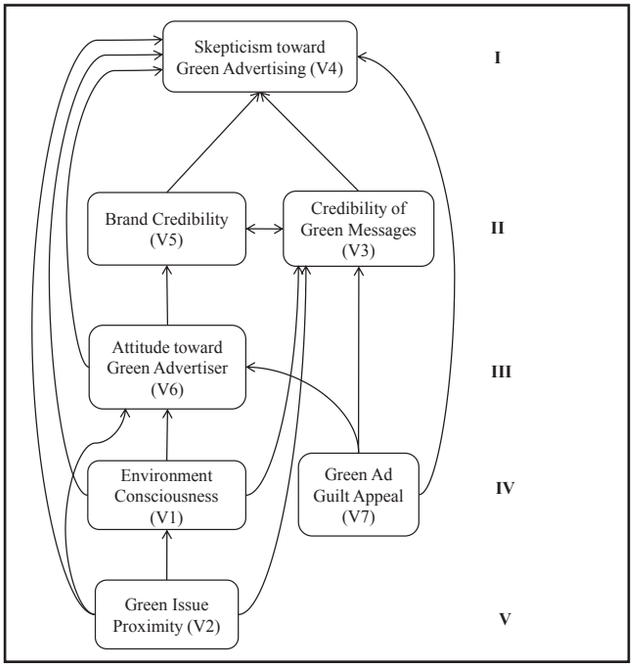


Fig. 1: ISM Framework for Skepticism toward Green Advertising

Subsequently, MICMAC analysis (Duperrin & Godet, 1973) was carried out to classify variables into four clusters (Figure 2) with an objective to analyze the driving power and the dependence of the respective variables. The taxonomy for the said clusters consists of the autonomous variables, the dependent variables, the linkage variables and the independent variables respectively. This clustering was done on the basis of reachability matrix (Table 4) wherein the driving power and the dependence of each of the respective variables had been identified.

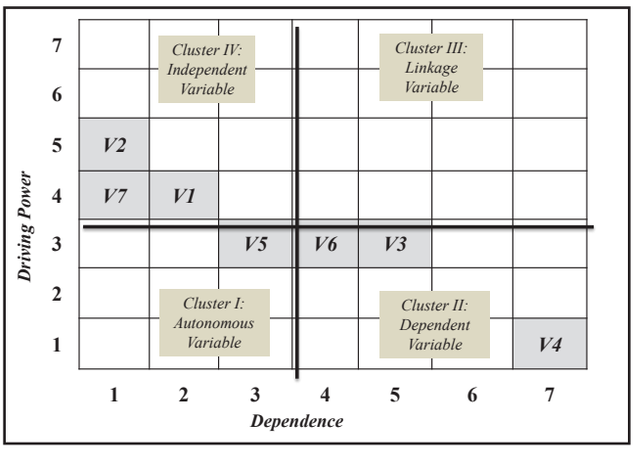


Fig. 2: MICMAC Analysis

The first cluster of autonomous variable(s) is characterized with weak dependence as well as weak driving powers. These variables though are relatively disconnected from the said phenomenon but may have few links, sometimes strong one. In the present study, variable Brand Credibility (V5) was found to be classified as autonomous variable. The second cluster of dependent variable(s), characterized with weak driving power but strong dependence, includes the variables Credibility of Green Messages (V3), Skepticism toward Green Advertising (V4) and Attitude toward Green Advertiser (V6). No variable was found to be part of third cluster of linkage variable(s) which has characteristics of strong driving power along with strong dependence. The fourth and last cluster consisting of independent variables signifies strong driving power and weak dependence thus same is often termed as key variable(s) (Ravi & Shankar, 2005). This cluster includes variables Environment Consciousness (V1), Green Issue Proximity (V2) and Green Ad Guilt Appeal (V7).

There is a possibility of various antecedents leading to skepticism toward green advertising which further may affect individuals' behavioral dispositions particularly the consumption of products supposedly green or environment friendly. The ISM framework and the subsequent MICMAC analysis suggest various propositions though those are subject to empirical investigation.

The study suggests that environment consciousness, green issue proximity and green ad guilt appeal can primarily be identified as the variables directly as well as indirectly leading to skepticism toward green advertising since these are the variables falling in the cluster of independent/key variable(s) (Figure 2), which is characterized with strong driving power but weak dependence. Thus the proposition can be made that the level of environment consciousness may have positive and significant impact on the degree of skepticism toward green advertising (Zinkhan & Carlson, 1995; Phau & Ong, 2007; Chang, 2012; do Paço & Reis, 2012; Bickart & Ruth, 2013; Matthes & Wonneberger, 2014). Environment consciousness can further be hypothesized as driver forming attitude toward green advertiser (Zinkhan & Carlson, 1995) and affecting credibility of green messages (Phau & Ong, 2007). The attitude toward green advertiser is further hypothesized to be driven by green issue proximity and green Ad guilt appeal (Cotte et al., 2005; Chang, 2012) while credibility of green messages is hypothesized to be driven by green

issue proximity and green Ad guilt appeal (Cotte et al., 2005; Chang, 2012) as well as brand credibility (Phau & Ong, 2007). The variables skepticism toward green advertising, credibility of green messages and attitude toward green advertiser are identified as dependent variables since the MICMAC analysis classifies them into second cluster of dependent variables with weak driving power but strong dependence. Here skepticism toward green advertising might be getting influenced by the said credibility of green messages (Cotte et al., 2005) and attitude toward green advertiser (Cotte et al., 2005).

The ISM framework further suggests that proximity to green issue(s) can be hypothesized as a factor influencing not only the degree of skepticism toward green advertising (Chang, 2012) but also the level of environment consciousness, hence it can be treated as moderating variable in the said proposition. The guilt appeal in green Ad might form favorable or unfavorable attitude toward the said green advertiser and subsequently might affect the degree of skepticism toward green advertising (Chang, 2012).

Further the cluster of autonomous variable(s) with the characteristics of weak dependence as well as weak driving powers can be hypothesized as mediating variable(s) affecting skepticism toward green advertising. Thus variable brand credibility is hypothesized to be driven by attitude toward green advertiser and also influencing the degree of skepticism toward green advertising. Here brand credibility and credibility of green messages may be treated as co-variant as both are hypothesized to be influencing each other (Phau & Ong, 2007).

Conclusion and Research Implications

The study is exploratory in nature and thus does not conclude. The interpretive structural modeling employed in the study, results a framework which provides a conceptual understanding about the possible factors influencing the degree of skepticism toward green advertising. However this is primarily a theoretical framework which needs to be tested empirically. Based on the ISM framework and subsequent MICMAC analysis various propositions have been formulated. This further requires operationalization of each of those variables identified and suggested in the framework and same can be subjected to empirical investigation. Multiple regression analysis or structural equation modeling (SEM) can be used to validate the framework. However SEM seems to be more feasible in this regard since the framework incorporates certain latent variables and also comprises of variables with dual nature, that is, being antecedent as well as consequence at the same time, thus the prepositions will be tested as a standard two step approach of SEM. It is a powerful multivariate technique which is used to establish the linear relationship between different variables of the study and is particularly useful in testing theories that contain multiple equations involving dependence relationships. It helps to identify direct and indirect effects in a complex system of variables, and allows including the mediating variables in the analysis easily. It provides a method of dealing with multiple relationships simultaneously and comprehensively for determining the goodness of fit measure of the sequential model (Bentler, 1990; Hair et al., 2007).

References

Agarwal, A., Shankar, R., & Tiwari, M. K. (2007). Modeling agility of supply chain. *Industrial Marketing Management*, 36(4), 443-457.

Banerjee, S., Charles, S. G., & Easwar, I. (1995). Shades of green: A multidimensional analysis of environmental advertising. *Journal of Advertising,* 24(2), 21-32.

Basil, D. Z., Ridgway, N. M., & Basil, M. D. (2006). Guilt appeal: The mediating effect of responsibility. *Psychology and Marketing,* 23(12), 1035-1054.

Basil, D. Z., Ridgway, N. M.,& Basil, M. D. (2008). Guilt and giving: A process model of empathy and efficacy. *Psychology and Marketing,* 25(1), 1-23.

Beltramini, R. F., & Stafford, E. R. (1993). Comprehension and perceived believability of seals of approval information in advertising. *Journal of Advertising*, 22(3), 3-13.

Bickart, B. A., & Ruth, J. A. (2012). Green eco-seals and advertising persuasion. *Journal of Advertising*, 41(4), 51-67.

Block, L. G. (2005). Self-referenced fear and guilt appeals: the moderating role of self-construal. *Journal of Applied Social Psychology,* 35(11), 2290-2309.

Bolanos, R., Fontela, E., Nenclares, A., & Pastor, P. (2005). Using interpretive structural modelling in strategic decision-making groups. *Management Decision*, 43(6), 877-895.

Boush, D. M., Friestad, M., & Rose, G. M. (1994). Adolescent skepticism toward TV advertising and

knowledge of advertiser tactics. *Journal of consumer research*, 165-175.

Branthwaite, A. (2002). Investigating the power of imagery in marketing communication: evidence-based techniques. *Qualitative Market Research: An International Journal*, *5*(3), 164-171.

Buda, R., & Zhang, Y. (2000). Consumer product evaluation: The interactive effect of message framing, presentation order and source credibility. *The Journal of Product & Brand Management,* 9(4), 229.

Calfee, J. E., & Ringold, D. J. (1994). The 70% majority: Enduring consumer beliefs about advertising. *Journal of Public Policy & Marketing*, 228-238.

Carlson, L., Grove, S. J., & Kangun, N. (1993). A content analysis of environmental advertising claims: a matrix method approach. *Journal of Advertising,* 22(3), 27-39.

Chan, R. Y. K. (2000). The effectiveness of environmental advertising: The role of claim type and the source country green image. *International Journal of Advertising,* 19(3), 349-375.

Chang, C. T. (2011). Guilt appeals in cause-related marketing: The subversive roles of product type and donation magnitude. *International Journal of Advertising,* 30(4), 587-617.

Chang, C. -T. (2012). Are guilt appeals a panacea in green advertising? The right formula of issue proximity and environmental consciousness. *International Journal of Advertising,* 31(4), 741-771.

Connolly, J.,& Prothero, A. (2003). Sustainable consumption: Consumption, consumers and the commodity discourse. *Consumption Markets and Culture,* 6(4), 275-91.

Cotte, J., Coulter, R. A., & Moore, M. (2005). Enhancing or disrupting guilt: The role of ad credibility and perceived manipulative intent. *Journal of Business Research*, 58(3), 361-368.

Coulter, R. H., & Pinto, M. B. (1995). Guilt appeals in advertising: What are their effects? *Journal of Applied Psychology,* 80(6), 697-705.

D'Souza, C., & Taghian, M. (2005). Green advertising effects on attitude and choice of advertising themes. *Asia Pacific Journal of Marketing and Logistics,* 17(3), 51-66.

do Paço, A. M. F., & Reis, R. (2012). Factors affecting skepticism toward green advertising. *Journal of Advertising*, 41(4), 147-155.

Duperrin, J. C., & Godet, M. (1973). *Methode de hierarchisation des elements d'un systeme*, Rapport Economique du CEA, Paris.

Edell, J. A., & Burke, M. C. (1987). The power of feelings in understanding advertising effects. *Journal of Consumer research*, 421-433.

Erdem, T., & Swait, J. (2004). Brand credibility, brand consideration, and choice. *Journal of Consumer Research*, 31(1), 191-198.

Farris, D. R., & Sage, A. P. (1975). On the use of interpretive structural modeling for worth assessment. *Computers & Electrical Engineering*, 2(2), 149-174.

Festinger, L. (1962). *A Theory of Cognitive Dissonance*, 2. Stanford, CA: Stanford university press.

Ford, G. T., Smith, D. B., & Swasy, J. L. (1990). Consumer skepticism of advertising claims: Testing hypotheses from economics of information. *Journal of Consumer Research*, 433-441.

Foreh, M. R., & Grier, S. (2003). When is honesty the best policy? The effect of stated company intent on consumer skepticism. *Journal of Consumer Psychology*, 13(3), 349-356.

Friestad, M., & Wright, P. (1995). Persuasion knowledge: Lay people's and researchers' beliefs about the psychology of advertising. *Journal of Consumer Research*, 62-74.

Ghingold, M. (1981). Guilt arousing communications: an unexplored variable, in K. Monroe (ed.) *Advances in Consumer Research,* 8. Ann Arbor, MI: Association for Consumer Research, 442-448.

Ghingold, M. (1981). Guilt arousing marketing communications: an unexplored variable. *Advances in consumer research*, 8(1), 442-448.

Ginsberg, J. M., & Bloom, P. (2004). Choosing the right green marketing strategy. *MIT Sloan Management Review,* 46 (1).

Goldsmith, E. R., Lafferty, A. B., & Newell, J. S. (2000). The impact of corporate credibility and celebrity credibility on consumer reaction to advertisements and brands. *Journal of Advertising,* 29(3), 43-54.

Grace, D., & O'Cass, A. (2002). Brand associations: Looking through the eye of the beholder. *Qualitative Market Research: An International Journal*, 5(2), 96-111.

Gregory, G. D., & Leo, M. D. (2003). Repeated behavior and environmental psychology: the role of personal involvement and habit formation in explaining water consumption1. *Journal of Applied Social Psychology*, 33(6), 1261-1296.

Hardesty, D. M., Carlson, J. P., & Bearden, W. (2002). Brand familiarity and invoice price effects on consumer evaluations: the moderating role of skepticism toward advertising. *Journal of advertising*, 31(2), 1-15.

Hartmann, P., & Apaolaza-Ibáñez, V. (2009). Green advertising revisited: conditioning virtual nature experiences. *International Journal of Advertising,* 28(4), 715-739.

Hasan, M. A., Shankar, R., & Sarkis, J. (2007). A study of barriers to agile manufacturing. *International Journal of Agile Systems and Management*, 2(1), 1-22.

Hawthorne, R. W., & Sage, A. P. (1975). On applications of interpretive structural modeling to higher education program planning. *Socio-Economic Planning Sciences*, 9(1), 31-43.

Hibbert, S., Smith, A., Davies, A., & Ireland, F. (2007). Guilt appeals: persuasion knowledge and charitable giving. *Psychology and Marketing,* 24(8), 723-742.

Hill, R. P., & Moran, N. (2011). Social marketing meets interactive media. *International Journal of Advertising,* 30(5), 815-838.

Huang, J. J., Tzeng, G. H., & Ong, C. S. (2005). Multidimensional data in multidimensional scaling using the analytic network process. *Pattern Recognition Letters*, 26(6), 755-767.

Huhmann, B. A., & Brotherton, T. P. (1997). A content analysis of guilt appeals in popular magazine advertisements. *Journal of Advertising,* 26(2), 35-46.

Kanter, D. L., & Wortzel, L. H. (1985). Cynicism and alienation as marketing considerations: Some new ways to approach the female consumer. *Journal of Consumer Marketing*, 2(1), 5-15.

Karna, J., Juslin, H., Ahoven, V., & Hansen, E. (2001). Green advertising: Greenwash or a true reflection of marketing strategies? *Journal of Corporate Environmental Strategy and Practice,* 33, 59-70.

Kim, H. S., & Damhorst, M. L. (1999). Environmental attitude and commitment in relation to ad message credibility. *Journal of Fashion Marketing & Management,* 3(1), 1-30.

Kollmuss, A., & Agyeman, J. (2002). Mind the gap: why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research*, 8(3), 239-260.

LaBarge, M. C., & Godek, J. (2006). Mothers, food, love and career the four major guilt groups? The differential effects of guilt appeals, in C. Pechmann & L. Price (eds) *Advances in Consumer Research,* Vol. 33, p. 511. Duluth, MN: Association for Consumer Research.

Laroche, M., Bergeron, J., & Barbaro-Forleo, G. (2001). Targeting consumers who are willing to pay more for environmentally friendly products. *Journal of Consumer Marketing,* 18(6), 503-521.

Latané, B. (1981). The psychology of social impact. *American Psychologist,* 36(4), 343-356.

Latané, B., & Bourgeois, M. J. (2001). Successfully simulating dynamic social impact: three levels of prediction, in J.P. Forgas & K.D. Williams (eds) *Social Influence: Direct and Indirect Process*. New York: Taylor & Francis, 61-76.

Latane, B. I. B. B., & Bourgeois, M. J. (2001). Successfully simulating dynamic social impact. *Social influence: Direct and indirect processes*, 3, 61.

Lutz, R. J. (1985). Affective and cognitive antecedents of attitude toward the ad: A conceptual framework. *Psychological processes and advertising effects*, 45-63.

MacKenzie, S. B., & Lutz, R. J. (1989). An empirical examination of the structural antecedents of attitude toward the ad in an advertising pretesting context. *The Journal of Marketing*, 48-65.

Mandal, A., & Deshmukh, S. G. (1994). Vendor selection using interpretive structural modelling (ISM). *International Journal of Operations & Production Management*, 14(6), 52-59.

Mangleburg, T. F., & Bristol, T. (1998). Socialization and adolescents' skepticism toward advertising. *Journal of Advertising*, 27(3), 11-21.

Manrai, L. A., Manrai, A. K., Lascu, D. N., & Ryans Jr, J. K. (1997). How green-claim strength and country disposition affect product evaluation and company image. *Psychology and Marketing,* 14(August), 511-537.

Marshall, M. E., & Mayer, D. W. (1992). Environmental training: It's good business. *Business Horizons*, 35(2), 54-57.

Matthes, J., & Wonneberger, A. (2014). The skeptical green consumer revisited: Testing the relationship between green consumerism and skepticism toward advertising. *Journal of Advertising*, 43(2), 115-127.

Mayer, R. N., Scammon, D. L., & Zick, C. D. (1992, May). *Turning the competition green: the regulation of environmental claims.* In Proceedings of the 1992 Marketing and Public Policy Conference (152-165). American Marketing Association Chicago, IL.

Mohr, L. A., Eroglu, D., & Ellen, P. S. (1998). The development and testing of a measure of skepticism toward environmental claims in marketers' communications. *Journal of Consumer Affairs,* 32(1), 30-55.

Nimon, W., & Beghin, J. (1999). Are eco labels valuable? Evidence from the apparel industry. *American Journal of Agricultural Economics,* 81(4), 801-11.

Obermiller, C. (1995). The baby is sick/the baby is well: a test of environmental communication appeals. *Journal of Advertising,* 24(2), 55-72.

Obermiller, C., & Spangenberg, E. R. (1998). Development of a scale to measure consumer skepticism toward advertising. *Journal of Consumer Psychology*, 7(2), 159-186.

Obermiller, C., Spangenberg, E., & MacLachlan, D. L. (2005). Ad skepticism: The consequences of disbelief. *Journal of Advertising*, 34(3), 7-17.

Peattie, K., & Crane, A. (2005). Green marketing: Legend, myth, farce or prophesy? *Qualitative Market Research: An International Journal,* 8(4), 357-370.

Pfohl, H. C., Gallus, P., & Thomas, D. (2011). Interpretive structural modeling of supply chain risks. *International Journal of physical distribution & logistics management*, 41(9), 839-859.

Phau, I., & Ong, D. (2007). An investigation of the effects of environmental claims in promotional messages for clothing brands. *Marketing Intelligence & Planning,* 25(7), 772-788.

Polonsky, M. J., & Rosenberger, P. J. III (2001). Reevaluating green marketing: a strategic approach. *Business Horizons,* September-October, 21-30.

Ratchford, B. T. (1987). New Insights about the FCB Grid. *Journal of Advertising Research*, 11(4), 24-38.

Ravi, V., & Shankar, R. (2005). Analysis of interactions among the barriers of reverse logistics. *Technological Forecasting and Social Change*, 72(8), 1011-1029.

Roberts, J. A. (1996). Green consumers in the 1990s: profile and implications for advertising. *Journal of business research*, 36(3), 217-231.

Rossiter, J. R., Percy, L., & Donovan, R. J. (1991). A better advertising planning grid. *Journal of Advertising Research*, 31(5), 11-21.

Royne, M. B., Martinez, J., Oakley, J., & Fox, A. K. (2012). The effectiveness of benefit type and price endings in green advertising. *Journal of Advertising*, 41(4), 85-102.

Sage, A. P. (1977). *Systems engineering: Methodology & applications*. IEEE Computer Society Press.

Saxena, J. P., & Vrat, P. (1992). Scenario building: A critical study of energy conservation in the Indian cement industry. *Technological Forecasting and Social Change*, 41(2), 121-146.

Schlegelmilch, B. B., Diamantopoulos, A., & Bohlen, G. M. (1994). *The value of socio-demographic characteristics for predicting environmental consciousness.* In Marketing Theory and Applications: The Proceedings of the 1994 American Marketing Associations Winter Educators Conference (Vol. 5, 348-349).

Schuhwerk, M. E., & Lefkoff-Hagius, R. (1995). Green or non-green? Does type of appeal matter when advertising a green product? *Journal of advertising*,24(2), 45-54.

Sherif, C. W., Sherif, M., & Nebergall, R. E. (1965). *Attitude and attitude change: The social judgment-involvement approach* (127-167). Philadelphia: Saunders.

Shrum, L. J., McCarty, J. A., & Lowrey, T. M. (1995). Buyer characteristics of the green consumer and their implications for advertising strategy. *Journal of Advertising,* 24(2), 71-82.

Singh, M. D., Shankar, R., Narain, R., & Agarwal, A. (2003). An interpretive structural modeling of knowledge management in engineering industries. *Journal of Advance Management Research*, 1(1), 28-40.

Srivastava, V., & Singh, T. (2010). Value creation through relationship closeness. *Journal of Strategic Marketing*, 18(1), 3-17.

Szymecko, L. (2003), Risk Communication. Retrieved from www.envirotools.org/presentations.Shtml, accessed between July-August 2014.

Takács-Sánta, A. (2007). Barriers to environmental concern. *Research in Human Ecology,* 14(1), 26-38.

Takács-Sánta, A. (2007). Barriers to environmental concern. *Human Ecology Review*, 14(1), 26.

Thakkar, J., Deshmukh, S. G., Gupta, A. D., & Shankar, R. (2006). Development of a balanced scorecard: an integrated approach of interpretive structural modeling (ISM) and analytic network process (ANP). *International Journal of Productivity and Performance Management*, 56(1), 25-59.

The GfK Roper Consulting. (2012). Green Gauge studies. Retrieved from http://www.scjohnson.com/Libraries/Download_Documents/SCJ_and_GfK_Roper_Green_Gauge.sflb.ashx accessed between July-August, 2014.

Thorgersen, J. (2002). Promoting 'green' consumer behavior with eco-labels, in Dietz, T. and Stern, P.C. (Eds), *New Tools for Environmental Protection,* Washington, DC: National Academy Press.

Turner, M., Xie, X., Lanmm, E., & Southard, B. (2009). *Encouraging mothers to get a mammogram: A cross-cultural examination of guilt appeals.* Paper presented at the annual meeting of the International Communication Association, New York.

Warfield, J. N. (1974). Developing subsystem matrices in structural modeling. *Systems, Man and Cybernetics, IEEE Transactions on*, (1), 74-80.

Warfield, J. N. (1976). *Societal systems: Planning, policy, and complexity*. New York: Wiley.

Watson, R. H. (1978). Interpretive structural modeling - A useful tool for technology assessment? *Technological Forecasting and Social Change*, 11(2), 165-185.

Zelezny, L. C., & Schultz, P. (2000). Psychology of promoting environmentalism: promoting environmentalism. *Journal of Social Issues*, 56(3), 365-371.

Zheng, Y. (2009). Association analysis on pro-environmental behaviors and environmental consciousness in main cities of East Asia. *Behaviormetrika*, 37(1), 55-69.

Zinkhan, G. M., & Carlson, L. (1995). Green advertising and the reluctant consumer. *Journal of Advertising*, 24(2), 1-6.

Article can be accessed online at http://www.publishingindia.com

An Application of Structural Equation Modelling to Determine the Inclusion of Climate Change Topics in MBA Education

Purba Halady Rao*, Rahul Pulupudi, Suman Sen***,**

Abstract

In the years to come India would be vulnerable to severe and unavoidable impacts of Climate Change such as Floods & droughts, water pollution and the associated health hazards, prevalence of diseases, the adverse impact on crop production, less availability of food and potable water, heat stress, mass migration, mortality, morbidity etc. To address these impacts of climate change, various strategies, recommendations, and action plans have been discussed across international agencies, government and non-government organizations. All the same the challenge is huge and it is believed that private sector could be equipped and inspired to help in meeting climate change challenges. In this context we have conducted this research to determine if inclusion of climate change education in MBA curriculum would indeed inspire, encourage and equip managers in the private sector to take up the climate change challenge to participate in helping communities survive and build up resilience to adapt to climate change.

Keyword: Sustainability, Climate Change, MBA, Education, Empirical, Structural Equation Modeling

Introduction

Why Managers need to Know about Climate Change Impacts

In the years to come India will become highly vulnerable to the impacts of climate change. The impacts are varied, unavoidable and severe such as Floods & droughts, water pollution and the associated health hazards, prevalence of diseases, the adverse impact on crop production, less availability of food and potable water, heat stress, mass migration, mortality, morbidity, and the impact on the quality of life (Climate Change Synthesis Report, WMO, UNEP 2001: 35-98). Thanks to research, conferences, media, and international agencies, the general awareness about climate change, its causes, and its adverse effects are widely known in today's world. To address these impacts of climate change, various strategies, recommendations, and action plans have been discussed across international agencies such as the UNFCC, the United Nations Environment Program, multilateral organizations such as the World Bank, the Asian Development Bank, the African Development Bank, and other national and international organizations (Asian Development Bank, Climate Change ADB Programs 2007: 2009). Even though the challenge is huge but addressing the challenge is very much needed.

Thus, given the magnitude of the help required to address climate change impacts, to set up the infrastructure to protect populations from floods as well as health-care systems all over the country, to invest in agriculture, water availability, sanitation facilities, etc.it might be beneficial to involve the private sector, in addition to government support which is present always, in addressing to help build the capacity for the communities affected by such calamities to survive the adversities caused by climate change

Corporate managers throughout the industry have been effective in planning and carrying out successful projects to bring about efficiency from the organizational perspective. It is expected and hoped that if they take it upon themselves in addressing impending climactic disaster, they can be equally effective too, this time with the goal of enhancing resilience and minimizing vulnerability. (http://unfccc.int/adaptation/workstreams/nairobi_work_programme/items/4623.php).

* Great Lakes Institute of Management,Dr. Bala V Balachandar Campus, East Coast Road, Manamai, Tamil Nadu, India. Email: purbarao@yahoo.in

** CEO & Founder,Company: Afterthought Feedback Services Pvt Ltd. City: Hyderabad, Country: India Email: rahulp@studynsurvey.com

*** CFO & Co-Founder, Afterthought Feedback Services Pvt Ltd, Bengaluru, Karanataka, India. Email: suman@studysurvey.com

Thus in the face of any kind of disasters and calamities it is believed that private sector can be catalyzed in helping through their involvement in the wider adaptation and building up resilience of communities. The unique expertise of the private sector, its capacity to innovate and produce new technologies for adaptation, and its financial leverage can form an important part of the multi-sectoral partnership that is required between governmental, private and non-governmental participants. (http://unfccc.int/adaptation/workstreams/nairobi_work_programme/items/4623.php).

In addition to helping communities cope with calamities, corporations could also make their own businesses safe and resilient. During the Chennai floods many auto parts manufacturing units got totally destroyed. If these companies could foresee climate change calamities coming, they could have thought of adaptation measures and built their units at higher grounds.

There is also another reason why private sector should be involved to address climate change(cc) related initiatives. The private sector through its operations generate huge amount of green-house gases that contribute to global warming/climate change. The total greenhouse gas (GHG) emissions from a selection of global 500 companies approximately amount to that of the USA and the EU15 combined. Not only do corporations have a significant climate footprint, but the impact of climate change on the business landscape is already noticeable.(Patenaude, 2011).

Through their manufacturing set up using coal and other fossil fuels, unlimited use of resources like water and electricity, using huge industrial air conditioning units that generate GHGs, non- renewable lighting, non-environment friendly transportation & distribution systems, corporations in the private sector have been major cause of global warming and climate change, whose effects can be felt today. Thus it is of utmost importance that private sector organizations get exposed to the perils of their activities and thereby learn how to use industrial activities which minimize the generation of GHGs and do not further enhance the effects of climate change any more.

The countries that negotiated the Paris deal now have a responsibility not only to talk about addressing climate change, but to help communities recover from the devastation it has already caused and will continue to cause, and enable them to learn how to adapt to future climate impacts and be climate resilient, and minimizing vulnerability.

Climate Change Concerns in Today's World Mitigation and Adaptation Strategies to Address Them

As is well known, increases in the concentration of GHGs in earth's atmosphere have given rise to global warming with anticipated impacts that are disastrous. These adverse effects of climate change are going to threaten economic growth in many regions like, health and well-being standards of millions of people, food security, availability of clean water and many other aspects of millennium development goals such as poverty alleviation. In order to address the impending crisis associated with the climate change worldwide forums have come up with strategies to implement and combat the impacts of climate change. The strategies thus designed can broadly be categorized as mitigation type strategies, preventing or minimizing the generation of GHGs which cause the climate change, and adaptation types which help communities to adapt to the inevitable impacts of climate change.

Initial initiatives at dealing with the problem of global warming focused on mitigation, that strives to aim at reducing and possibly stabilizing the GHG concentrations in the atmosphere (UNFCCC 1992).However, the increase of GHGs which are already there would continue to cause disastrous impacts which are unavoidable. Thus there is no alternative but to learn to adapt to such impacts and try and reduce the damage to the extent possible, leading to the development of adaptation strategies. In fact, what happened in Chennai required a thorough preparedness system for adaptation of the communities to the unavoidable disasters descending upon them. It is here that the adaptation strategy to help communities adapt/ survive due to the effects of climate change was most appropriate.

These two strategies, mitigation (trying to reduce generation of GHGs which cause climate change) and adaptation-coping with the effects which cannot be avoided anymore), are intricately linked - the more we mitigate, the less we have to adapt (Adaptation to Climate Change, GLCA, 2009).

In today's world, ever since awareness of climate change started growing and policy makers developed initiatives

in addressing the climate change phenomenon, the focus was mostly on mitigation and hardly on adaptation, though it was realized that adaptation was urgently needed too.(Schipper, 2006; Tol, 2005, Klein et al., 2007). Also throughout history, mitigation and adaptation were regarded as two fundamentally different approaches to the same problem,ignoring possible synergies and trade-offs between them.

However, recently in research and in general literature the potentials of combining both approaches have been considered within the scientific community (Burch and Robinson 2007, Dowlatabadi, 2007, Goklany 2007, Jones, Dettmann, Park, Rogers, and White 2007, Klein et al. 2007, and Swart and Raes, 2007). Also, it has emerged that they need to be integrated because they both influence each other… the more one mitigates, lesser is the need for adaptation.

Mitigation Strategy

There have been extensive research, seminar discussions, conferences, and international forums on building awareness of mitigation strategies with the objective of reducing the generation of GHGs. (Climate Change ADB Programs, 2007, Joint MDB Report 2008, Munasinghe, 2008).

Some of such mitigation initiatives are as follows: Please see Table 1.

Table 1: Some Common Mitigation Initiatives at the Individual Level

Preventing/protesting against trees to be cut down, because trees absorb CO2, a green -house gas
Planting trees,
Using solar lighting system, solar energy is a form of clean energy
Using energy efficient lighting, less generation of GHG
Switching off lights when additional light is not needed in the room,
Using LED based decorative lighting, LED lighting generates less GHG
Using system to capture methane gas from landfills,so that methane does not go to atmosphere
Using emission free cars,
Reducing car trips and reducing consumption of gasoline,
etc.

(reference: Climate change, ADB Programs… Strengthening Mitigation and Adaptation in Asia and the Pacific, 2007)

In addition there are mitigation initiatives at the organization level. These are as follows.

Please see Table 2.

Table 2: Mitigation Possibilities at Organization Level

Using climate change friendly production system and cooling system, which do not generate GHGs.
Use input materials which do not produce GHGs.
Using cleaning agents which do not generate GHGs.
Using boilers which do not produce GHGs
Ensuring that buildings have adequate insulation in order to minimize the energy needed for cooling in summer.
Ensuring that natural light is used where possible in order to reduce demand for lighting.
Using solar energy in production,
Producing products which do not generate GHGs upon use by customer.
Etc.

(reference: Climate change, ADB Programs…Strengthening Mitigation and Adaptation in Asia and the Pacific, 2007).

Adaptation Strategy

Adaptation in the context of climate change refers to how communities, groups of people, sectors, regions or even countries develop systems to better cope with the impacts of cc. (Brooks, 2005, Smit et al. 2003, Pielke, 2007) Also in the climate context adaptation is defined as the adjustments in individual groups and institutional behavior in order to reduce society's vulnerability to climate. The adaptation process can be anticipatory and/or reactive, and could be autonomous or planned (Fankhauser et al., 1999; Smit et al., 2000).

Adaptation strategies are often specific to local conditions and are locally relevant action plans aimed at reducing climate-related risks. These action plans are community-based development projects that are meant to promote information sharing, develop early warning systems (to warn of an impending calamity) and preparedness plans, develop more diverse crop strains that can withstand a variety of conditions (heat, drought, salt, etc.), bolster social capital and resilience, increase storage capacity for fresh water by building reservoirs or by recharging aquifers, improve public health infrastructure, and bolster disease surveillance. These strategies would be valuable regardless of the exact impacts of climate change at a particular time or location.

(GEF 2009).

Adaptation strategies are therefore evolved or should be developed looking at the exact nature of vulnerability of the target communities, who are in danger of climate change impacts and suited to help them best adapt to address the impacts… totally specific to local conditions (Kelly and Adger, 2000; Downing, 2001; Turner et al., 2003; Smit and Pilifosova, 2003; Yohe et al., 2003; Adger, 2006).

Involvement of Corporate Managers to Address Climate Change Concerns

As observed by India last year, when the Chennai flooding happened in Nov/Dec 2015, the city was totally caught unaware as to what to do and were unprepared.

Chennai has always been strongly affected by hazards related to climate change, due to proximity to the coastal zone and the increase of heavy rainfall events during the monsoon season. There have been initiatives to address flooding because this hazard arises most often and affects many people in Chennai. Also it has been noted that resilience could be improved by strengthening adaptation measures, especially in the slum areas that were considered to be the most vulnerable because of the greatest exposure to climate change problems, least resilient and lack of proper drainage system etc. (Potarazu, Sreedhar, 2015), http://www.cnn.com/2015/12/19/opinions/ potarazu-chennai-flooding.

All the same, when the Nov 2015 flooding calamity happened and continued for few weeks, it was felt that if corporate organizations were involved and organized to address such issues, the miseries of communities could have been widely lessened. Also during the floods huge number of manufacturing units of SME category were washed off and endured huge losses in monetary terms. Much of this could have been avoided if proper know-how was available in advance. (Urban resilience to adaptation to climate change in Chennai, 2014).

In the face of this uphill task of combating climate change, the paper proposes to explore if it would be feasible to involve the private sector, over and above governmental and NGO initiatives, in working towards mitigation and adaptation strategies to reduce the generation of GHGs and also help build the capacity for the vulnerable poor to survive the adversities caused by climate change. (Owens, 2000, Pidgeon et al., 2003; Norton and Leaman, 2004). So far, the involvement of the private sector in the climate change issue has primarily been only in the realm of mitigation strategies and carbon trading, where a company reduces the emission level, equates the reduction to what is called a certified emissions reduction (CER), and sells it to countries that need CER. However, though emission saving is achieved at the world level, carbon trading does not necessarily and directly lead to adaptation or reduction in vulnerability for the poor, which might be an urgent need in today's world.(McCarthy et al., 2001).

Also, after policies regarding mitigation and adaptation have been designed, it would be necessary to translate the policies into action plans, implement them in real terms and bring about actual changes in abatement in climate change. (Halady and Rao, 2010). Preclude to this would be enhancing the awareness to the impacts of climate change as well as to initiatives which do exist to address the challenge (Burgess et al., 1998).

For the private sector, especially for small and medium enterprises (SMEs), it is very important to know what exactly to do in the face of cc calamities…how to make their own businesses resilient. While climate change poses a number of risks to vulnerable communities and businesses around the world, many opportunities are unfolding for private companies to implement actions towards reducing risks to their business operations, as well as investing in adaptation action in vulnerable regions in a sustainable and profitable manner.

Adaptation activities for corporate organizations may thus relate either to ensuring the resilience of business operations, or the provision of technologies or services that assist in the adaptation for vulnerable communities. (http://unfccc.int/adaptation/workstreams/nairobi_work_programme/items/6547.php).

However, a thorough knowledge base and awareness would be needed for building up the mindset of corporate managers,for them to take up planning and preparedness in effective and organized manner. It is believed that such awareness and knowhow could be imparted to the managers' right from the time they take up their educational degrees…perhaps from the MBA days.

The Role of MBAs and Business Schools

Higher education programs such as MBA should often have an obligation to create campus climate-action plans that address the curricular component of this problem. (Tare, M, 2016, http://www.triplepundit.com/2016/02/how-institutions-of-higher-education-can-address-climate-change/)

The Carbon Commitment (formerly the American College & University Presidents' Climate Commitment or ACUPCC) is such a "high-visibility effort" to address climate change by creating a network of colleges and universities that have committed to neutralize their greenhouse gas emissions and accelerate the research and educational efforts of higher education to equip society to re-stabilize the earth's climate.

The Carbon Commitment seeks to create connections with higher educational institutions in order to carry out two goals: to make an agreement with these colleges and universities that they will commit to eliminate their net greenhouse gas emissions from specified campus operations and to focus on education and the institutions' ability to promote research of sustainability programs and empower the "higher education sector to educate students, create solutions, and provide leadership-by-example for the rest of society".

Being in positions of leadership MBAs have often had the opportunities to not only make decisions that affect entire organizations, but also inspire and mentor others. However, reviews have found that most MBA programs have not seriously addressed the issue of climate change yet in their curricula (Whiteway, Parker, 2013, http://www.educationpost.com.hk/resources/mba/160308-mba-career-help-five-strategic-imperatives-for-digital-brand-building). This needs to be corrected on urgent and immediate basis.

There is considerable influence of business schools on business practitioners. An important proportion of corporate leaders hold a degree in business administration or an MBA. Not only do corporations have a significant climate footprint, but the impact of climate change on the business landscape is already noticeable. Yet, meeting the managerial challenges that climate change brings requires knowledge that is only being imparted moderately in business education and scholarship today.

The knowledge acquired during business studies is also widely applied in practice. Strategic paradigms developed or taught by business scholars such as Porter's Five Forces (Porter, 1979), the Value Chain (Porter, 1985) and the SWOT analysis are being used ubiquitously in the corporate world. Yet, even acknowledging that the managerial challenges that climate change brings requires expert knowledge and it is only being moderately addressed in business education today. (Climate change diffusion: While the world tips, business schools lagPatenaude, G., Global Environmental Change 21, 2011).

Climate Change Topics in MBA Education in India

Outside India there are some universities that are starting to show an interest in teaching climate change topics in their MBA curricula .Some of these are termed green MBA programs such as Corporate Knights and Beyond Grey Pinstripes.But all the same, these are very few who strive to excel at teaching about the most pressing environmental challenge the world has ever faced, climate change, and how companies can profit by being part of the solution. Looking at executive MBAs, again there are hardly any such programs which cover these topics. http://www.triplepundit.com/2011/03/top-5-executive-education-programs-climate-change/.

Within India Teri University offers academic programs at the Masters level, including, MBA and Doctoral programs on Sustainability and Climate Change. (http://www.teriuniversity.ac.in/).

Their curriculum has the objective to offer MBA program combining traditional MBA structure and topics with climate change and sustainability challenge. They believe that such a curriculum to enable their MBAs would have the strategic leadership to become holistic and competent business leaders with long term perspective which will work alongside the global perspective in the future. (http://www.teriin.org/mba-admission/?utm_source=yellobar_teriu_mba2016&utm_medium=yellobar&utm_campaign=TERIU_mba_yellobar2016#programmes).

To help our country and to help Indians develop resilience against climate change calamities MBAs can take advantage of their business leadership position to make climate change and sustainability a company priority…. both for the betterment of their own operations (protecting against calamities) and reach out and help communities around them. They can use the mastery of strategizing and decision making, which they have acquired in their MBA study, to (a) look within the walls of the factory to climate-change-proof their own operations and (b) to look beyond the walls of the factory to help communities adapt and survive the disaster impacts of climate change. This leads to our research question.

Research Question

The current research explores what categories of Climate Change topics can be included in the MBA curriculum to make aspiring business leaders aware of possible actions which exist for them to address climate change and also to inspire them take up such initiatives to address climate change.

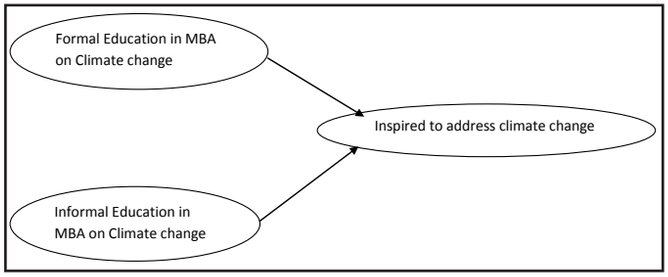


Fig. 1: Research question: Do Formal and Informal Education in MBA Lead to Inspire Corporate Managers to Address Climate Change Impacts… both in Mitigation as Well as in Adaptation Context

The Research Methodology

In the absence of extensive literature and published documents on the topic of MBA education and climate change strategies, it was decided to use a research instrument in the form of a questionnaire and get responses from corporate managers as to their agreement on what categories of topics on climate change could have and should have been included in their MBA curriculum… so that they would have been aware of what initiatives they may pursue, to address climate change and they would have developed an inclination and inspiration to pursue the same.

Research Instrument… a Survey Questionnaire

It was proposed that the research instrument, the questionnaire, would present different climate change related topicsto the respondent and ask him/her how important that topic will be to be included in the MBA curriculum.(Formal education on Climate Change).

These topics would include following items as given in Table 3

Table 3: MBA Including Formal Education On

* reasons which caused climate change
* negative health impact of cc,
* rise of diseases and morbidity impact of cc,
* lack of food and clean water due to cc,
* impacts of cc such as glaciers melting, sea level rising etc.
* impacts of cc on poor, vulnerability of poor,
* international agreements on cc,
* how important is cc in the global scenario.
* different ways to reduce generation of GHGs,
* possible organizational initiatives to reduce GHGs,
* use electrical systems with reduced GHG generation,
* use less power, go for renewable energy,
* conserve electricity, water and other resources.,
* use transportation systems that reduce GHGs,
* use solar lighting systems etc.
* possible ways to help communities set up systems to address impending impacts of cc.
* possible ways to help communities combat prevalence of disease.

* water availability impacts due to climate change
* vulnerability of poor due to cc,
* possibility of initiatives to help communities adapt to impacts of cc,
* ways to build awareness to mitigation strategies,
* adaptation strategies such as developing early warning systems,
* adaptation strategies such as predicting Rainfall patterns,
* adaptation strategies such as helping communities in building resilience,
* adaptation strategies to Reducing vulnerability of affected communities,
* adaptation strategies such as developing crops which can withstand climatic hazards.
* adaptation strategies such as developing Risk management and risk reduction

In addition, it was also proposed to obtain importance ratings from the respondents on Informal education/ seminars on cc organized by MBA school, MBA school being a green campus, MBA school using solar energy for lighting and/or water heating, MBA School having bio-gas generation plant using kitchen waste etc., MBA School having waste water recycling plant.

Population, Sampling and Data Collection Method

The empirical research was conducted on a population of MBA students in their last term before graduation. They had work experience between 2-15 years, before they joined MBA. The MBA schools considered for the research were IIM Ahmedabad, Great Lakes Institute of Management, IIM Kolkata, Indian School of Business, IIT Mumbai and Myra School of Business. The survey questionnaire was digitalized by online survey portal called study & Survey. Com. The link was extended to the population of students in these MBA schools.

A total of 95 survey responses were received, with margin of error of 10%.

Developing constructs on formal education on Mitigation strategies, Adaptation strategies, Impacts of Climate Change, Climate Change relevance and Informal exposure/seminars on climate change.

In the questionnaire that was constituted as the research instrument, there were items on each of the above themes such as Mitigation strategies, Adaptation strategies, Impacts of Climate Change, Climate Change relevance and Informal exposure/seminars on climate change. The respondents were asked to give their importance ratings on a 4-point likert scale to each of them.

The constructs and the items (exact questions) which constituted them were:

Please rate the aspects of MBA education you feel are important to be included in the MBA curriculum.(Formal education on Climate Change)

Formal Education on Impacts of Climate Change

Formal education on negative health impact of cc,

rise of disease and morbidity impact of cc,

lack of food and clean water due to cc,

impacts of cc such as glaciers melting, sea level rising etc., vulnerability of the poor, and water non availability impacts of climate change.

Formal Education on Mitigation Strategies

Formal education on building awareness to different ways to reduce generation of GHGs individually, Organizational initiatives to reduce GHGs,

Initiatives to use electrical systems with reduced GHG generation,

use less power,

go for renewable energy,

Encourage organizations to conserve electricity, water and other resources,

Use transportation systems to reduce GHGs,

use solar lighting systems etc.

Formal Education on Adaptation Strategies

Formal education that there are ways to help communities set up systems to address impending impacts of cc. and combat prevalence of disease.

Initiatives to help communities adapt to impacts of cc,

Adaptation strategies such as developing early warning systems,

adaptation strategies such as predicting Rainfall patterns, helping communities to building resilience,

Reducing vulnerability of affected communities developing crops which can withstand climatic hazards, Developing Risk management and risk reduction

Formal Education on Relevance of Climate Change in todays' world

Formal education on international agreements on cc, Formal education on how important is cc in the global scenario.

Informal Education/Seminars in MBA School

Informal education/seminars on cc organized by MBA School

MBA School being a green campus equipped with solar energy, waste recycling etc. MBA School using solar energy for lighting and/or water heating MBA School having bio-gas generation plant using kitchen waste etc. MBA School having waste water recycling plant

Additional questions

In addition to asking respondents to give importance ratings to each of the above items under each construct, the respondents were also asked the following questions:

*During your MBA was any course offered on:

Environmental Sustainability

Causes and Impacts of global warming/climate change

*Did you take up such a course?

The respondents were also asked

*The Business School where you did your MBA did it have a

Green campus
Solar water heating system
Solar lighting system
Environment friendly waste recycling system
Water recycling system
Bio-gas generation system

Finally, the respondents were asked to what extent they were inspired to take up mitigation and adaptation initiatives as corporate managers.

Inspired to take up Initiatives to address Impacts of Climate Change.

*D1: How inspired are you to take up initiative to reduce generation of GHG (mitigation)?

*D2: How inspired are you to take up initiative to help communities set up systems to combat impacts of cc(adaptation)?

Structural Equation Modeling (SEM) to Address the Research Question

In order to explore the research question, if inclusion of climate change related topics in MBA curriculum would inspire them take up initiatives to address climate change, a linear SEM approach was used (Jo¨reskog and So¨rbom, 1993). This approach was applied to explore the causal relationships between the different latent constructs explained in the previous section, such as: Formal Education on Impacts of Climate Change, Formal Education on Mitigation Strategies, Formal Education on Adaptation Strategies, Formal Education on Relevance of Climate Change in todays' world, Informal education/ seminars in MBA School and Inspired to take up Initiatives to address Impacts of Climate Change.

SEM estimates a series of separate but interdependent multiple regression equations simultaneously. The research has drawn upon the theory and the research objectives to determine which independent variable will predict which dependent variable. The proposed relationships were then translated into a series of structural equations for each dependent variable.

The significance of the overall models was determined by the chi-square value, the corresponding degrees of freedom and the associated overall p-value with the significance of 0:05 which would be required to be more than 0.05.

The individual linkages between any two constructs were tested using the critical ratio, which would be required to be > 1.96 for significance at 5% level of significance.

The confirmatory factor analysis, CFA, under the general category of structural equation modeling, was used to validate the conceptual model involving the constructs using AMOS Graphics for Windows, estimating the regression weight of each link (arrow) and the associated significance.

The estimation procedure used was under the maximum likelihood estimation (MLE) procedure, which was known to provide valid results with sample sizes as small as 50. In addition to overall model p-value, the indicator defined as chi-square/degrees of freedom, Goodness of fit index (GFI), adjusted goodness of fit index (AGFI), and root mean square residual (RMSR) were additional indicators used to evaluate the validity of the model.

The Chi square/ degrees of freedom should be < 2 for a good fit.

Several sets of analyses were conducted which included iterations of sets of structural equation models that were run to test variations of the model with alternate paths to assess the importance of aspects of the conceptual model.

The model was designed using IBM-SPSS-AMOS Version 20.0.0(Build 788). The outcome of the structural equations among the latent constructs yielded the results as given below. Please see Table 4.

Results for Data Analysis by Structural Equation Modeling

Table 4: Measures of Goodness of Fit

Chi-square/degrees of freedom	=	*0.826*
Overall model p-value	=	0.936
GFI	=	.892
NFI	=	.893
CFI	=	1.000
AGFI	=	.840
RMSEA	=	0.000

The Maximum likelihood estimates are as given in Table 5

Table 5: Maximum Likelihood Estimates: Regression Weights: (Group number 1 - Default model)

			Estimate	*S.E.*	*C.R.*	*P*
formal education on CC impact	<---	informal education on CC	.532	.220	2.414	.016
formal education on adaptation	<---	Formal education on impact of climate change	.779	.183	4.251	***
formal education on adaptation	<---	informal education on CC	.536	.203	2.642	.008
formal education on mitigatation	<---	formal education on	.994	.198	5.009	***
formal education on world	<---	adaptation	.784	.188	4.165	***
Inspired to address CC	<---	formal impact education	-.294	.333	-.884	.377
Inspired to address CC	<---	formal world education	.171	.114	1.499	.134
Inspired to address CC	<---	informal education on CC	.307	.269	1.143	.253
Inspired to address CC	<---	formal education on mitigation	-.351	.201	-1.744	.081
Inspired to address CC	<---	formal education on adaptation	1.034	.488	2.118	.034

Based on the Maximum Liklihood Estimates, their significance levels and critical ratios, the following significant links emerge.

Significant Links in the Final SEM model as observed above

(1) Informal MBA education on Climate change → Formal MBA education on Climate Change Impacts.

(2) Formal MBA education on Climate Change Impacts → Formal MBA education on Adaptation strategies

(3) Formal MBA education on adaptation → Formal Education on Relevance of Climate Change in to-days' world

(4) Formal MBA education on adaptation strategies → Formal MBA education on Mitigation strategies

(5) Informal MBA education on Climate change → Formal MBA education on Adaptation strategies

(6) Formal MBA education on Adaptation strategies → Inspired to take up Initiatives to address Impacts of Climate Change.

From the results it emerges that Informal education in the MBA school regarding the campus being equipped with conservation initiatives, climate change and renewable energy initiatives as well as seminars/workshops organized on climate change, lead to formal education on climate change impacts being included in the curriculum.

The informal system of education also leads to formal education on adaptation initiatives being included in the curriculum, which again leads to Relevance of Climate Change Education in Today's world.

The only construct which directly leads to MBAs being inspired to address climate change concerns is formal MBA education on adaptation which however is impacted by Informal MBA education on Climate change.

The results therefore indicate that in order to get the private sector/corporate managers inspired to take up initiatives to address climate change, MBA education should include adaptation topics in the MBA curriculum, which has direct link to them taking up such initiatives. At the same time, one observes the indirect link which informal education on climate change has on inspiring managers to take up initiatives to address climate change… and encourage MBA schools to also organize seminars/workshops on climate change in the campus. Also this informal education would essentially require campus being climate change friendly, use conservation systems, waste recycling systems, bio gas generation systems and renewable energy such as use of solar energy wherever possible.

Hence the link to private sector/corporate managers get inspired to take up initiatives to address climate change, starts with informal education that would essentially require campus being climate change friendly, etc. This finding is of tremendous value because it tells all MBA campuses to start becoming Climate Change & Sustainability friendly that will inspire outgoing managers to be inspired.

References

Adger, W. N. (2006). Vulnerability. *Global Environmental Change,* 16(3), 268-281.

Asian Development Bank. (2007). Climate change ADB programs: Strengthening mitigation and adaptation in Asia and the Pacific.

Brooks, N., Adger, W. N., & Kelly, P. M. (2005). The determinants of vulnerability andadaptive capacity at the national level and the implications for adaptation.*Global Environ. Change,* 15, 151–163.

Burch, S., & Robinson, J. (2007). A framework for explaining the links betweencapacity and action in response to global climate change. *Climate Policy,* 7(4), 304-316.

Dowlatabadi, H. (2007). On integration of policies for climate and global change. *Mitigation and Adaptation Strategies for Global Change,* 12(5), 651-663.

Downing, T. E. (2001). Climate Change Vulnerability: Linking Impacts andAdaptation. Report to the GoverningCouncilofthe UnitedNationsEnvironment Programme. Environmental Change Institute, Oxford, UK

GEF. (2009). *Facilitating an international agreement on climate change: Adaptation to climate change*.

Fankhauser, S., Smith, J. B., & Tol, R. S. J., 1999. Weathering climate change:some simple rules to guide adaptation decisions. *Ecological Economics,* 30, 67-78.

Goklany, I. M. (2007). Integrated strategies to reduce vulnerability and advanceadaptation, mitigation, and sustainable development. *Mitigation and Adaptation Strategies for Global Change,* 12(5), 755-786.

Halady, I., & Rao, P. (2010).Does awareness to climate change lead to behavioral change?, *International Journal of Climate Change, Strategies and Management* (UK: Emerald).

Jones, R. N., Dettmann, P., Park, G., Rogers, M., & White, T. (2007). The relationship

between adaptation and mitigation in managing climate change risks:a regional response from North Central Victoria, Australia. *Mitigation and Adaptation Strategies for Global Change,* 12(5), 685–712.

Kelly, P. M., & Adger, W. N. (2000). Theory and practice in assessingvulnerability to climate change and facilitating adaptation. *Climate Change,* 47, 325-352.

McCarthy, J., Canziani, O. F., Leary, N. A., Dokken, D. J., & White, K. S. (2001*). Climate Change*: *Impacts, Adaptations and Vulnerability in Developing Economies,* World Bank, Washington, DC.

Munasinghe, M. (2008). *Addressing the Sustainable Development and Climate Change Challenges Together: Applying the Sustainomics Framework.* Procedia Social and Behavioral Sciences, 41(2010), 6634-6640.

Norton, A., & Leaman, J. (2004). *The day after tomorrow: Public opinion on climate change.*

MORI Social Research Institute, London.

Owens, S. (2000). Engaging the public: Information and deliberation in environmental policy.

Environment & Planning A, 32, 1141-1148.

Pidgeon, N., Kasperson, R. E., & Slovic, P. (2003). *The social amplification of risk,* Cambridge University Press, Cambridge

Pielke, R. A., Prins, G., Rayner, S., & Sarewitz, D. (2007). Climate change 2007: Lifting the taboo on adaptation. *Nature,* 445(7128), 597–-598.

Porter, M. E. (1979). How competitive forces shape strategy. *Harvard Business Review,* 57, 137-145.

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance.* Free Press, New York.

Potarazu, S.(2015). Chennai floods a climate change wake-up call for world. Retrieved from http://edition.cnn.com/2015/12/19/opinions/potarazu-chennai-flooding/index.html

Schipper, E. L. F. (2006). Conceptual history of adaptation in the UNFCCC process. *Review of European Community & International Environmental Law,* 15(1), 82-92.

Smit, B., Burton, I., Klein, R. J. T., & Wandel, J. (2000). An anatomy of adaptation toclimate change and variability. *Climatic Change,* 45(1), 223-251.

Smit, B., & Pilifosova, O. (2003). From adaptation to adaptive capacity and vulnerability reduction. In: Smith, J.B., Klein, R.J.T., Huq, S. (Eds.), *Climate Change, Adaptive Capacity and Development.* Imperial College Press, London., UK, 51-70.

Smit, B., & Wandel, J. (2006). Adaptation, adaptive capacity and vulnerability. *Global Environ. Change,* 16, 282-292.

Swart, R. J., & Raes, F. (2007). Making integration of adaptation and mitigation work:

Mainstreaming into sustainable development policies? *Climate Policy,* 7(4), 288-303.

Tare, M. (2016). *Climate Change and Sustainability in the curriculum. Institute for Sustainability and Global Impact at the University of Texas at Arlington*

Tol, R. S. J. (2005). Adaptation and mitigation: Trade-offs in substance and methods. *Environmental Science and Policy,* 8(6), 572-578.

Turner, B. L., Kasperson, R. E., Matson, P. A., McCarthy, J. J., Corell, R. W., Christensen, L., Eckley, N., Kasperson, J. X., Luers, A., Martello, M. L., Polsky, C., Pulsipher, A., & Schiller, A. (2003). *A framework for vulnerbility analysis in sustainability science.* Proceedings of the National Academy of Sciences 100, 8074-8079.

United Nations Climate Change Conference.(2009). Retrieved from https://en.wikipedia.org/wiki/2009_United_Nations_Climate_Change_Conference.

UNFCCC. (1992). *Poverty and climate change: Reducing the vulnerability of the poor.*A contribution to the eighth conference of the Parties to the United Nations Framework Convention on Climate Change.

Yohe, G., Strzepek, K., Pau, T., & Yohe, C. (2003). Assessing Vulnerability in the context of changing socioeconomic conditions: A study of Egypt. In: Smith, J.B., Klein, R.J.T., Huq, S. (Eds.), *Climate Change, Adaptive Capacity and Development.* Imperial College Press, London.

Article can be accessed online at http://www.publishingindia.com

Distribution of Traffic Accident Times in India - Some Insights using Circular Data Analysis

Arnab Kumar Laha*, Pravida Raja A. C.*, Dilip Kumar Ghosh**

Abstract

Traffic accidents are a major hazard for travellers on Indian roads. These are caused by a variety of reasons including the bad condition of roads, traffic density, lack of proper training of drivers, slack in enforcement of traffic rules, poor road lighting etc. It is further known that certain times of the day are more prone to traffic accidents than others. In this paper we investigate the distribution of traffic accident times using the data published annually by the National Crime Records Bureau (NCRB) over the period 2001-2014 using the tools of circular data analysis. It is seen that the observed distribution of the traffic accident times in most years is bimodal. Thus, several modelling strategies for bimodal distributions are tried which include fitting of mixture of von-Mises distributions and mixture of Kato-Jones distribution. It is seen from this analysis that the distribution of the traffic accident times are changing over the years. Notably, the proportion of accidents happening in late night has reduced over the years while the same has increased for late evening hours. Some more insights obtained from this analysis are also discussed.

Keyword: Circular Statistics, Kato-Jones Distribution, Mixture Distribution, Traffic Accidents, Von-Mises Distribution

Introduction

Increasing incidents of road traffic accidents pose a major societal problem in India and other developing countries. According to Aderamo (2012), road traffic accidents are decreasing in developed countries and increasing in developing nations. Many researchers have paid attention in determining the factors that significantly affect injury. It is mentioned in research by David and Hyder (2006) that the application of policies and interventions to control traffic accidents can decrease the societal cost. Petrol rationing, an improvement in traffic enforcement, setting up of speed bumps, legislation and the enforcement of the use of helmets for cyclists and motorcyclists are examples of such interventions. There are many factors which can increase the risk of traffic accidents such as construction and maintenance of roads and vehicles, driver's behaviour, speed of vehicle, highway characteristics, traffic characteristics, and weather condition. Cools, Moons, and Wets (2010) focussed on the effect of weather conditions on daily traffic intensities (the number of cars passing a specific segment of a road) in Belgium and the results of their analysis indicates that snowfall, rainfall and wind speed reduces the traffic density but high temperature increases the traffic density. Statistical modelling for predicting road accidents is gaining popularity in the literature on road safety. Kong, Lekawa, Navarro, McGrath, Cohen, Margulies, & Hiatt (1996) studied bicyclist accidents in China and Germany during 2001 to 2006 and the analysis shows there were similarities and differences between the two countries especially for the frequency, age distribution of the fatalities and the road environment where accidents occurred. The paper also suggests the importance of the usage of helmet and improvement of road environment for reduction of accidents and fatalities in China.

The time of day has an important role in traffic accidents. It is believed that even though the traffic density is less in the night compared to the day time, the number of accidents is more in the night time. According to studies, reduced visibility is an important contributor to the night time traffic accidents. Owens and Sivak (1993) studied the role of reduced visibility in night time road fatalities recorded by the U.S Fatal Accident Reporting Systems from 1980 through 1990.

* Indian Institute of Management, Ahmedabad, Gujarat, India.Email: arnab@iima.ac.in
** Saurashtra University, Rajkot, Gujarat, India.

Plainis, Murray, and Pallikaris (2006) compared the road injury data under dim and bright conditions for two EU countries and showed that low luminance is likely to contribute to the disproportionate number of road traffic injuries occurring at night. According to them, the presence of road lighting leads to substantial decrease in the severity of injuries in both countries, despite the fact that they have dramatically different injury rates.

Data

The yearly data for the period 2001-2014 are obtained from the reports entitled Accidental Deaths and Suicides in India (ADSI) published by National Crime Records Bureau, India. The data consist of number of traffic accidents at different times of the day at national level in 53 Indian cities. The given data are grouped in 3 hourly time intervals 0 – 3am, 3 – 6 am, 6 – 9 am, 9-12 noon, 12-3 pm, 3-6 pm, 6-9 pm and 9-12 midnight aggregated over the years. The snapshot of the final data retrieved from http://ncrb.gov.inis given in Table 1.

Table 1: Snapshot of Data. Each Figure Indicates the Number of Accidents in the Specified Time Interval for a given Year

Time of occurrence	*2001*	*2002*	*2008-2012*			*2013*	*2014*
0-3	23869	23894	-	-	-	28332	26068
3-6	30949	28485	-	-	-	35385	32554
6-9	41612	41110	-	-	-	52771	52279
9-12	50293	51904	-	-	-	67224	69042
12-15	47291	48794	-	-	-	65974	68918
15-18	49925	52026	-	-	-	73141	77830
18-21	46663	48287	-	-	-	74411	76334
21-24	33118	34934	-	-	-	45763	47873
Total	323720	329434	-	-	-	443001	450898

We map each accident time in 24-hour period onto a point on the unit circle i.e., an angle between 0 to 2π radians. Every 45 degrees (π/4 radians) on the circle denotes 3 hours in real time. 03:00 a.m. is mapped to 0 degree on the circle. The histogram of the time of accident occurrence for the years 2001, 2008 and 2014 are given in the Fig. 1.

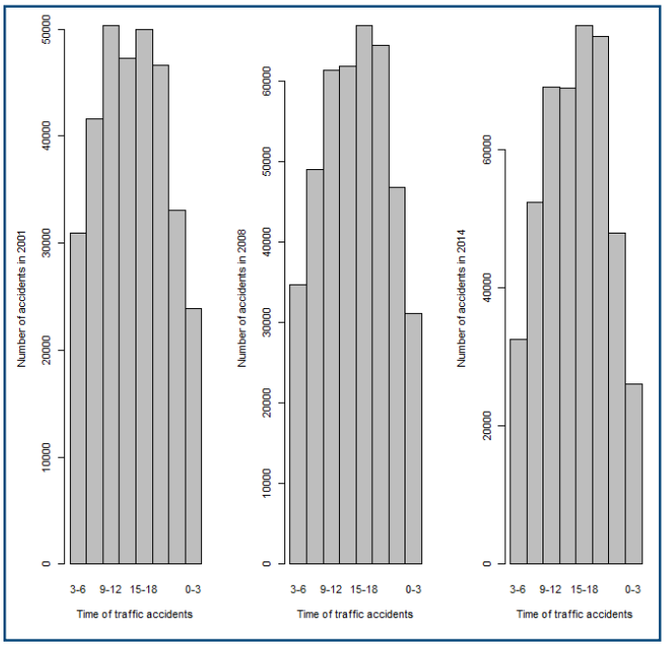


Fig. 1: Histogram of the Time of Traffic Accidents for the Years 2001, 2008 and 2014

Circular Data

Circular data are data measured in angles and occur in a variety of fields. They are commonly summarised as locations on a unit circle or as angles over a 360°or 2π radians range. Angular data arise in two ways, natural angles and observations which can be converted to angles. In this paper, we are mainly interested in the time of the day when the accident occurred. The time of accident is converted to an angle in the following manner. Let x be the time of the day recorded in hours. Then the corresponding angle would be $\frac{x}{24} * 2\pi$ radian. For example, if an accident occurred at 4 a.m., then the corresponding angle would be $\frac{4}{24} * 2\pi = \frac{\pi}{3}$ radians.

The purpose of this work is to analyse Indian road traffic accident times data using circular data analysis. Recently several authors have used Circular Statistics to analyse and model distributions of random variables that are cyclic in nature. Brunsdon and Corcoran (2006) used circular statistics to analyse time patterns in crime incidence. They analysed a data set related to the reports of criminal damage in the city of Cardiff, Wales during the period July 1999 to June 2001. The circular plot of the data (see Fig. 4 of Brunsdon & Corcoran, 2006) shows a bimodality in which the frequency of reporting crimes peaks around 11 PM and 10 AM rounded to the nearest hour. Faggian, Corcoran, and McCann (2013) introduced the use of circular statistics to study the interregional graduate migration flows in Britain. Corcoran, Chhetri, and Stimson (2009) applied circular statistics to analyse journey to work data. They calculated the direction and frequency of each journey using bespoke tools developed in a Geographic Information System (GIS) environment. They used the geographical angle of journey from an origin to a destination as a central variable for analysis which is circadian in nature (see Fig. 1 of Corcoran *et al*., 2009). The application of circular statistics in particular circular mean direction of travel and circular spread gives an indication of the modality direction of the commuter from any movements given origin zone. Gill and Hangartner (2010) studied an interesting application of circular data in political science. They developed a circular regression model for terrorism events.

Circular Distributions

The most popular circular distributions used in applied work is the von-Mises (vM) or Circular Normal distribution (CN), which is described in sub-section below. However, there are many alternative circular distributions and a comprehensive account of the properties of these distributions can be found in Mardia and Jupp (2000) and Jammalamadaka and SenGupta (2001). Recently an extension of the Circular Normal distribution known as Kato-Jones distributions is finding increasing use in applied work. We discuss these distributions in sub-sections below.

Circular Normal (von-Mises) Distribution

Circular Normal (CN) distribution plays a central role in the analysis of circular data. This distribution was introduced as a statistical model by von-Mises (1918). A circular random variable Θ is said to have a CN distribution with mean direction parameter μ and concentration parameter κ if it has the probability density function (p.d.f.)

$$f(\theta;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa\cos(\theta-\mu)),$$

$$0 \le \theta < 2\pi,\ 0 \le \mu < 2\pi, \kappa > 0$$

Where $I_0(.)$ is the modified Bessel function of order 0. This distribution is symmetric about μ and unimodal. We will denote this distribution as $CN(\mu,\kappa)$. Another interesting property of $CN(\mu,\kappa)$ distribution is that, for sufficiently large κ, the CN distribution can be approximated by a linear normal distribution with mean μ and variance $\frac{1}{\sqrt{\kappa}}$.

Kato-Jones (KJ) Distributions

Kato and Jones (2010) proposed a family of four parameter distributions on the circle that contains von-Mises and wrapped Cauchy distributions as special cases. This family of distributions is derived by transforming von-Mises distribution through Mobius transformation. The density function of this distribution is

$$f(\theta;\ \mu,\ \nu,\ r,\ \kappa) = \frac{1-r^2}{2\pi\ I_0(\kappa)} \exp\left\{\frac{\kappa\left(\xi\cos(\theta-\eta)-2r\cos\nu\right)}{1+r^2-2r\cos(\theta-\gamma)}\right\}$$

$$\times \frac{1}{1+r^2-2r\cos(\theta-\gamma)};\ 0 \le \theta < 2\pi$$

where $\gamma = \mu + \nu,\ \xi = \sqrt{r^4 + 2r^2\cos 2\nu + 1}$ and $\eta = \mu + \arg\left\{r^2\cos 2\nu + 1 + ir^2\sin 2\nu\right\}$

such that $0 \le \mu, \nu < 2\pi$, $\kappa > 0$, and $0 \le r < 1$.

A three parameter family of distributions can be derived as a special case when $\nu = 0$ or $\nu = \pi$. In this case the above four parameter distribution reduces to

$$f(\theta;\mu,\kappa,r) = \frac{1-r^2}{2\pi\, I_0(\kappa)} \exp\left\{\frac{\kappa\left(1+r^2\right)\cos(\theta-\mu) - 2r}{1+r^2-2r\cos(\theta-\mu)}\right\}$$

$$\times \frac{1}{1+r^2-2r\cos(\theta-\mu)};\ 0 \le \theta < 2\pi \quad .$$

The distribution is symmetric about $\theta = \mu$ and $\mu + \pi$ and is unimodal when $0 \le r < 1$. The parameter μ is the directional mean. Symbolically, we will write $\theta \sim KJ(\mu,\ \kappa, \mathrm{r})$. The above model involves the von-Mises $(r = 0)$, wrapped Cauchy $(\kappa = 0)$ and uniform distributions $(\kappa = r = 0)$ as special cases. As $\kappa \to \infty$, the Kato-Jones distribution tends to $\mathrm{N}(\mu, \omega_{\mathrm{r}})$ where the standard deviation $\omega_{\mathrm{r}} = \frac{1-\mathrm{r}}{1+\mathrm{r}}$.

Finite Mixture of Distributions

Finite mixtures of distributions (FM) has seen many applications in the linear data context. Some applications in the circular data context have also been reported in the literature. A random variable X is said to follow a k-component mixture distribution of densities $f_1, f_2, ..., f_k$ if its p.d.f. is of the form

$$p(x) = \sum_{j=1}^{k} \pi_j f_j(x)$$

where π_j s is a set of probabilities also known as mixing weights such that $\sum_{j=1}^{k} \pi_j = 1$ and $f_j(x)$, $\mathrm{j} = 1,2,...,\ \mathrm{k}$ are the component densities.An up-to-date brief overview of the developments in FM models can be seen in Zhang and Huang (2015).Roy et al. (2012) designed mixture model based colour image segmentation in the LCH colour space using a Circular- Linear distribution. Mooney, Helms, and Jollife (2003) analysed Sudden Infant Death Syndrome (SIDS) data for the UK from 1983 to 1998 and in their study, they pointed out that for some years, there seems to be more than one mode. Later Jiang (2009) analysed this data set by fitting a von-Mises distribution and a mixture of two von-Mises distributions and reported that for most years the data could be fitted using a mixture of two von-Mises distributions. Jiang (2009) also analysed the fatal traffic crash time data in the United States in 2007 and showed that for Washington and the District of Columbia a mixture of two von-Mises distributions fitted the dataset.

Modelling Traffic Accident Times

In this paper we model the time of traffic accidents data using a mixture of two circular distributions. We consider two models

(1) a two component mixture of Circular Normal distributions $\alpha\ CN(\mu_1,\kappa_1) + (1-\alpha)\ CN(\mu_2,\kappa_2)$ and

(2) a two component mixture of Kato and Jones distributions $\alpha\ KJ(\mu_1,\kappa_1,r) + (1-\alpha)\ KJ(\mu_2,\kappa_2,r)$.

In both these cases we restrict μ_1 in $[0,\pi)$ and μ_2 in $[\pi, 2\pi)$. Let $f_{CN}(\theta;\mu_1,\mu_2,\kappa_1,\kappa_2,\alpha)$ be the pdf of mixture of Circular Normal distributions $\alpha\ CN(\mu_1,\kappa_1) + (1-\alpha)\ CN(\mu_2,\kappa_2)$. We define

$$p_1 = \int_0^{\frac{\pi}{4}} f_{CN}(\theta)d\theta \quad p_2 = \int_{\frac{\pi}{4}}^{\frac{\pi}{2}} f_{CN}(\theta)d\theta \quad , \ldots$$

$$p_8 = \int_{\frac{7\pi}{4}}^{2\pi} f_{CN}(\theta)d\theta \ . \qquad (1)$$

Note that p_{j}'s(j=1,…,8) depend on the unknown parameters $\mu_1,\mu_2,\kappa_1,\kappa_2,\alpha$. Now to estimate the unknown parameters we apply the minimum chi-square method (Berkson, 1980) which is briefly discussed below.

Let $x_1, x_2, ..., x_n$ be the given data. Define $e_j = np_j$ and $\mathrm{n}_{\mathrm{j}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \varepsilon_{\mathrm{ij}}$ where

$$\varepsilon_{ij} = \begin{cases} 1 & if \quad \mathrm{x_i} \in \left[(j-1)\frac{\pi}{4},\ j\frac{\pi}{4}\right) \\ 0 & otherwise \end{cases}$$

for i=1,…,n and j=1,…,8. Consider the function $g(\mu_1,\mu_2,\kappa_1,\kappa_2,\alpha) = \sum_{j=1}^{8} \frac{(n_j - e_j)^2}{e_j}$. To obtain

the estimates of the parameters we minimize the function g subject to the conditions $0 \le \mu_1, \mu_2 < 2\pi$, $\kappa_1 > 0,\ \kappa_2 > 0,\ 0 < \alpha < 1$. Since this function g is difficult to minimize analytically, we adopt a direct numerical minimisation approach using the function DEOptim in R (Mullen, Ardia, Gil, Windover, and Cline, 2011).Fig. 2(a)–(c) show estimated parameters of mixture of von-Mises distribution.

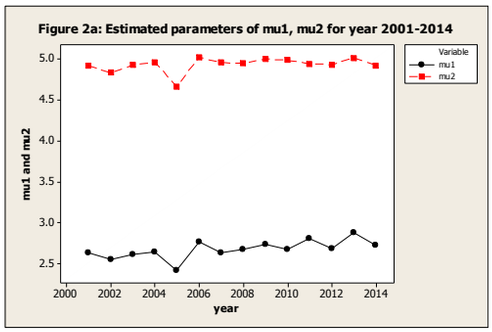


Fig. 2 (a): Estimated parameters of mu1, mu2 for year 2001-2014

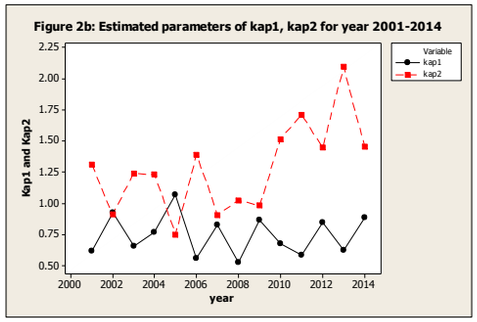


Fig. 2 (b): Estimated parameters of kap1, kap2 for year 2001-2014

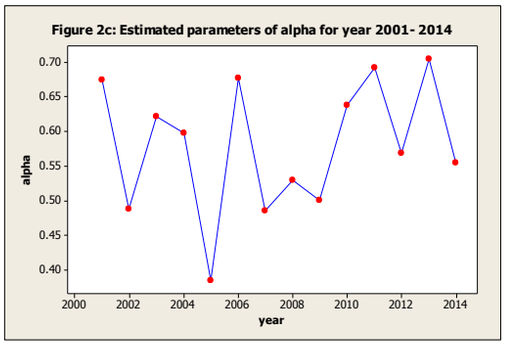


Fig. 2 (c): Estimated parameters of alpha for year 2001-2014

Let $f_{KJ}(\theta;\mu_1,\mu_2,\kappa_1,\kappa_2,r,\alpha)$ be the pdf of mixture of Kato-Jones distributions $\alpha\ \mathrm{KJ}(\mu_1,\kappa_1,r)+(1-\alpha)\ KJ(\mu_2,\kappa_2,r)$.We define p_j's (j=1,…, 8) in a manner analogous to (1) and obtain the minimum chi-square estimates of the parameters $(\mu_1,\mu_2,\kappa_1,\kappa_2,r,\alpha)$ using direct numerical minimisation of g subject to the conditions $0 \le \mu_1, \mu_2 < 2\pi$, $\kappa_1 > 0, \kappa_2 > 0,\ 0 \le r < 1, 0 < \alpha < 1$.Fig. 3(a)–(d) show estimated parameters of mixture of Kato-Jones distributions for year 2001 to 2014.

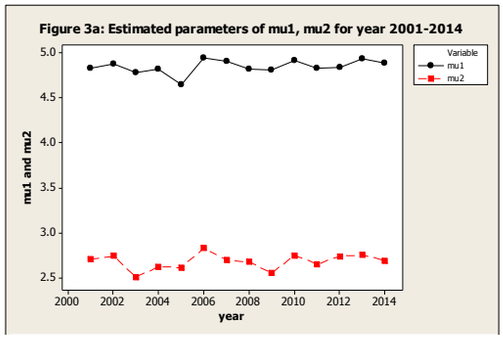


Fig. 3 (a): Estimated parameters of mu1, mu2 for year 2001-2014

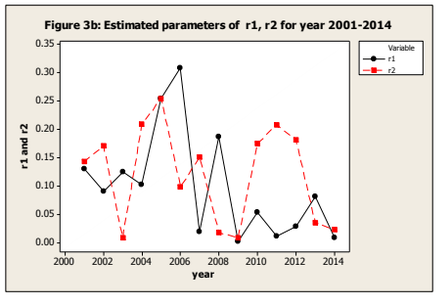


Fig. 3 (b): Estimated parameters of r1, r2 for year 2001-2014

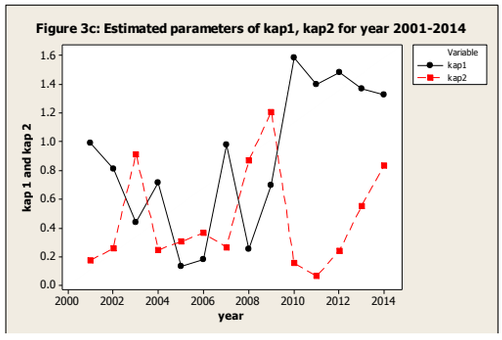


Fig. 3 (c): Estimated Parameters of kap1, kap2 for year 2001-2014

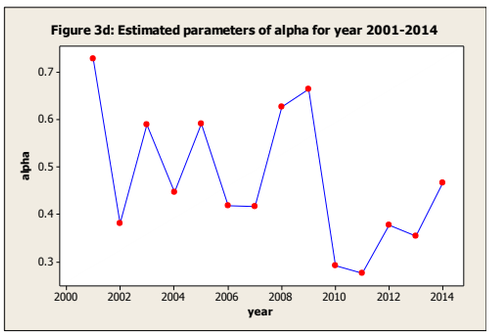


Fig. 3 (d): Estimated Parameters of alpha for year 2001-2014

Using these estimated parameters, we calculated the proportion of accidents happening in certain times of the day from year 2001-2014 which has been shown in Table 2.

Table 2: Proportion of Accidents Happening in Certain Times of the Day from Year 2001 to 2014 using Mixture of Circular Normal Distributions(CN) and Mixture of Kato-Jones (KJ) Distributions

		Time of occurrence							
Year	*Model*	*0-3*	*3-6*	*6-9*	*9-12*	*12-15*	*15-18*	*18-21*	*21-24*
2001	Actual	0.073	0.096	0.129	0.155	0.146	0.154	0.144	0.102
	vM	0.076	0.093	0.131	0.154	0.149	0.153	0.145	0.100
	KJ	0.079	0.093	0.128	0.153	0.147	0.155	0.147	0.097
2002	Actual	0.073	0.086	0.125	0.158	0.148	0.158	0.147	0.106
	vM	0.073	0.085	0.128	0.155	0.151	0.157	0.148	0.103
	KJ	0.076	0.084	0.122	0.158	0.151	0.155	0.149	0.103
2003	Actual	0.073	0.092	0.124	0.156	0.144	0.157	0.148	0.106
	vM	0.075	0.089	0.127	0.149	0.147	0.156	0.152	0.105
	KJ	0.077	0.090	0.128	0.149	0.146	0.158	0.149	0.102
2004	Actual	0.070	0.088	0.122	0.160	0.144	0.159	0.151	0.106
	vM	0.072	0.086	0.131	0.154	0.146	0.156	0.151	0.104
	KJ	0.075	0.084	0.125	0.158	0.146	0.158	0.152	0.103
2005	Actual	0.073	0.089	0.122	0.161	0.146	0.163	0.146	0.101
	vM	0.073	0.087	0.129	0.152	0.153	0.161	0.145	0.101
	KJ	0.076	0.082	0.123	0.159	0.149	0.165	0.147	0.098
2006	Actual	0.077	0.090	0.199	0.150	0.144	0.157	0.154	0.110
	vM	0.078	0.088	0.121	0.146	0.148	0.156	0.155	0.109
	KJ	0.080	0.087	0.118	0.149	0.147	0.154	0.157	0.106
2007	Actual	0.077	0.086	0.116	0.150	0.147	0.157	0.153	0.114
	vM	0.079	0.083	0.119	0.147	0.147	0.156	0.154	0.113
	KJ	0.078	0.085	0.118	0.149	0.150	0.151	0.156	0.112
2008	Actual	0.075	0.084	0.118	0.148	0.149	0.161	0.155	0.112
	vM	0.076	0.079	0.123	0.146	0.145	0.162	0.157	0.112
	KJ	0.079	0.084	0118	0.146	0.149	0.162	0.155	0.107
2009	Actual	0.069	0.083	0.115	0.152	0.146	0.163	0.152	0.120
	vM	0.075	0.078	0.116	0.149	0.152	0.159	0.158	0.114
	KJ	0.076	0.079	0.119	0.151	0.150	0.159	0.153	0.111

		Time of occurrence							
Year	Model	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24
2010	Actual	0.068	0.088	0.121	0.151	0.145	0.160	0.161	0.107
	vM	0.071	0.084	0.124	0.149	0.145	0.158	0.162	0.107
	KJ	0.073	0.084	0.119	0.153	0.146	0.157	0.165	0.103
2011	Actual	0.067	0.085	0.118	0.152	0.150	0.166	0.160	0.102
	vM	0.069	0.083	0.121	0.149	0.151	0.165	0.161	0.101
	KJ	0.076	0.093	0.131	0.154	0.149	0.152	0.145	0.100
2012	Actual	0.063	0.078	0.118	0.153	0.148	0.167	0.166	0.107
	vM	0.063	0.075	0.121	0.152	0.149	0.167	0.168	0.105
	KJ	0.067	0.075	0.114	0.151	0.149	0.172	0.169	0.102
2013	Actual	0.064	0.080	0.119	0.152	0.149	0.165	0.168	0.103
	vM	0.063	0.078	0.119	0.151	0.150	0.164	0.171	0.103
	KJ	0.067	0.080	0.119	0.148	0.146	0.166	0.172	0.102
2014	Actual	0.058	0.072	0.116	0.153	0.153	0.173	0.169	0.106
	vM	0.059	0.070	0.117	0.153	0.153	0.171	0.171	0.104
	KJ	0.061	0.068	0.114	0.154	0.155	0.169	0.172	0.106

It can be seen from Table 2 that the proportion of accidents happening in late night (9pm – 3am) has reduced over the years (0.175 in 2001 to 0.164 in 2014) while the same has increased for late evening hours (6-9pm) which has been captured by both the models under consideration. We use the Schwarz Information Criterion (SIC) (Schwarz,1978) to choose the best model among these two. The SIC is defined as $\text{SIC} = -2 \log L\left(\hat{\theta}\right) + k \ln(n)$, where $L\left(\hat{\theta}\right)$ is the likelihood function for the model evaluated at the estimated parameter value $\hat{\theta}$, k is the number of parameters and n is the sample size. The likelihood is calculated for the two models and the corresponding SIC year wise values are given in Table 3 for the years 2005 - 2009.

Table 3: SIC for Mixture of Von-Mises and Mixture of Kato-Jones Distributions from 2005 to 2009

Year	SIC (Mix vM)	SIC (Mix KJ)
2005	569.3544	365.9092
2006	277.066	258.6512
2007	252.944	232.7688
2008	186.6506	197.3286
2009	820.9194	823.3113

We see that in some of the cases the SIC is minimum for mixture of Kato-Jones distributions whereas in some other cases it is minimum for mixture of von-Mises distributions. Since the family of two component mixture of Kato-Jones distributions contains the family of two component mixture of von-Mises distributions, we consider the former for modelling the time of the accidents.

Change Point Problem

The change point problem is introduced in statistics by Page (1955) in the context of statistical quality control. It has been discussed quite extensively in the literature for linear data. Let $x_1, x_2, ..., x_n$ be independent observations. It is often of interest to know if there exists a(unknown) point s; $1 \le s \le n-1$ such that $x_1, x_2, ...x_s$ are independently and identically distributed (i.i.d) F_0 and $x_{s+1}, x_{s+2}, ...x_n$ are i.i.d. F_1 ($F_0 \neq F_1$). Here s is called the change point of the data. This formulation is usually referred to as at most one (or single) change point problem (AMOC). If $s = n$ then all observations are from F_0 or we say that there is no change point. In change point problem, one is interested to test

$H_0 : x_i's$ are i.i.d F_0 against the alternative

H1 : there exist s, $1 \le s \le n-1$, such that $x_1, x_2, \ldots, X_s$ are i.i.d. F_0 and $x_{s+1}, x_{s+2}, \ldots, x_n$ are i.i.d. F_1

Here F_i , i = 0,1 may be known or unknown and may contain one or more unknown parameters.

Lombard (1986) proposed rank based non-parametric procedures to test the presence of change point in a sequence of angular observations. Ghosh, Jammalamadaka, and Vasudaven (1999) considered a generalised likelihood ratio test procedure and a Bayes procedure for change-point problems of the mean direction of the CN distribution. Grabovsky and Horvath (2001) suggested a modified procedure to detect changes in circular data. Sengupta and Laha (2008a) introduced a likelihood integrated method for exploratory graphical analysis of change point problem with directional data. Sengupta and Laha (2008b) also discussed the problem of detecting change in the mean direction of the circular normal distribution when the concentration parameter is unknown using Bayesian analysis.

Chen and Gupta (1997) approached the change point problem as a model selection problem. Specifically they consider the models

M_0 : The accident time distribution is F_0 for all the years against

$\mathrm{M_s}$: The accident time distribution is F_0 for the first s years and is F_1 for the years $s+1$ to n.

They then propose to use SIC to choose the best model among these models. We assume F_0 belongs to the family of two-component mixture of Kato-Jones distribution $\alpha\ \mathrm{KJ}\left(\mu_{1b},\kappa_1,r\right)+\left(1-\alpha\right)\ KJ\ \left(\mu_{2b},\kappa_2,r\right)$, where μ_{1b} and μ_{2b} are the mean directions of the two component Kato-Jones distributions before change point occurred. We also assume that F_1 is a member of the above family but with different parameters. i.e., F_1 is $\alpha\ \mathrm{KJ}\left(\mu_{1a},\kappa_1,r\right)+\left(1-\alpha\right)\ KJ\ \left(\mu_{2a},\kappa_2,r\right)$ where μ_{1a} and μ_{2a} are the mean directions of the two component Kato-Jones distributions after the change point. Since there are unknown parameters in both F_0 and F_1 we only investigate the presence of change point in the period 2005 to 2009. Following Chen and Gupta (1997), we use the minimum SIC criterion for choosing the best model amongst $M_0,...,M_{14}$. To compute the SIC for $\mathrm{M_0}$ we need to compute the likelihood L_0. It is not difficult to observe $L_0=\prod_{s=1}^{14}\frac{n_s\ !}{\mathrm{n_{1s}}\,!n_{2s}\,!...n_{8s}\,!}p_1^{\mathrm{n_{1s}}}\,p_2^{\mathrm{n_{2s}}}...p_8^{\mathrm{n_{8s}}}$

where $p_1=\int_0^{\frac{\pi}{4}}f_{KJ_M}\left(\theta\right)d\theta$, $p_2=\int_{\frac{\pi}{4}}^{\frac{\pi}{2}}f_{KJ_M}\left(\theta\right)d\theta$, … , $p_8=\int_{\frac{7\pi}{4}}^{2\pi}f_{KJ_M}\left(\theta\right)d\theta$; n_{1s} is the number of accidents occurring between 3 - 6 am in years, $n_s=\sum_{j=1}^{8}n_{js}$ and $f_{KJ_M}\left(\theta;\mu_1,\mu_2,\kappa_1,\kappa_2,r,\alpha\right)$ be the pdf of mixture of Kato-Jones distributions $\alpha\ \mathrm{KJ}\left(\mu_1,\kappa_1,r\right)+\left(1-\alpha\right)\ KJ\left(\mu_2,\kappa_2,r\right)$. For computing SIC for $\mathrm{M_s}$, we need to compute the likelihood L_s which is given by $L_s=\prod_{k=1}^{s}\frac{n_k\ !}{\mathrm{n_{1k}}\,!n_{2k}\,!...n_{8k}\,!}p_{1b}^{n_{1k}}\,p_{2b}^{n_{2k}}...p_{8b}^{n_{8k}}$

$\prod_{k=s+1}^{14}\frac{n_k\ !}{\mathrm{n_{1k}}\,!n_{2k}\,!...n_{8k}\,!}p_{1a}^{n_{1k}}\,p_{2a}^{n_{2k}}...p_{8a}^{n_{8k}}$ where

$p_{1b}=\int_0^{\frac{\pi}{4}}f_{KJ_{b_M}}\left(\theta\right)d\theta$ $p_{2b}=\int_{\frac{\pi}{4}}^{\frac{\pi}{2}}f_{KJ_{b_M}}\left(\theta\right)d\theta$,…., $p_{8b}=\int_{\frac{7\pi}{4}}^{2\pi}f_{KJ_{b_M}}\left(\theta\right)d\theta$ and $p_{1a}=\int_0^{\frac{\pi}{4}}f_{KJ_{a_M}}\left(\theta\right)d\theta$, $p_{2a}=\int_{\frac{\pi}{4}}^{\frac{\pi}{2}}f_{KJ_{a_m}}\left(\theta\right)d\theta$,….., $p_{8a}=\int_{\frac{7\pi}{4}}^{2\pi}f_{KJ_{a_M}}\left(\theta\right)d\theta$

. Here $f_{KJ_{b_M}}\left(\theta;\mu_{1b},\mu_{2b},\kappa_1,\kappa_2,r,\alpha\right)$ and $f_{KJ_{a_M}}\left(\theta;\mu_{1a},\mu_{2a},\kappa_1,\kappa_2,r,\alpha\right)$ are the p.d.f. of mixture of Kato-Jones distributions before and after the change point respectively.

Table 4 gives SIC of the models we considered for year 2005-2009.

Table 4: Model for 2005-2009 and the corresponding SIC values

Model	SIC KJ
M_0	13282.21
M_5	11716.88
M_6	10755.03
M_7	13286.83
M_8	10743.92
M_9	10787.72

It can be seen from Table 4 that M_8 has the lowest SIC which indicates that there is a change point and the year of the change is 2008. The estimated parameters of the mixture of Kato-Jones distributions for the years 2001-2008 are μ_{1b} = 1.82 radian (104.19°), μ_{2b} = 4.13 radian (236.62°), κ_1 = 0.6, κ_2 = 1.42, r = 0.01 and α = 0.64, and the same for the years 2009 -2014 are μ_{1l} = 1.98 radian (113.68°), μ_{2a} = 4.16 radian (238.36°), κ_1 = 0.6, κ_2 = 1.42, r = 0.01 and α = 0.64. Thus it can be observed that the accident time during the years 2001-2008 has modes at 10:34am and 7:18pm, and the same during the years 2009-2014 are at 10:58am and 7:29pm.

Conclusion

The observed distribution of the traffic accident times in India for most years in the period 2001-14 is seen to be bimodal. This bimodal distribution has been modelled using a two component mixture of Kato-Jones distributions. The distribution is seen to be a decent fit to the observed data. It is further seen that the distribution of the traffic accident times are changing over the years. Notably, the proportion of accidents happening in late night has reduced over the years while the same has increased for late evening hours. A formal change point analysis indicates the presence of a change point in the year 2008.

References

Aderamo, A. J. (2012). Assessing the trends in road traffic accident casualties on Nigerian roads. *Journal of Social Sciences*, 31, 19-25.

Berkson, J. (1980). Minimum Chi-Square, not maximum likelihood. *Annals of Statistics*, 8(3) , 457-487.

Brunsdon, C., & Corcoran, J. (2006). Using circular statistics to analyse time patterns in crime incidence. *Computers, Environment and Urban Systems*, 30, 300-319.

Chen, J., & Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92, (438), 739-747.

Cools, M., Moons, E., & Wets, G. (2010). Assessing the impact of weather on traffic intensity. *American Meteorological Society, Weather, Climate and Society*, 2, 60-68.

Corcoran, J. , Chhetri, P., & Stimson, R. (2009). Using circular statistics to explore the geography of the journey to work, *Progress in Regional Science*, 88(1), 119-132.

David, M. B, & Hyder, A. A. (2006). Modelling the cost effectiveness of injury interventions in lower and middle income countries: Opportunities and challenges. *Cost Effectiveness and Resource Allocation,* 4, 2, 1-11.

Faggian, A., Corcoran, J., & McCann, P. (2013). Modelling geographical graduate job search using circular statistics, *Papers in regional Science*, 92(2), 329-343.

Ghosh, K., Jammalamadaka, S. R., & Vasudaven, M. (1999). Change-point problems for the von-Mises distribution. *Journal of applied statistics*, 26(4), 423-434.

Gill, J., & Hangartner, D. (2010). Circular data in Political Science and how to handle It, *Political Analysis*, 18 (3), 316-336.

Grabovsky, I., & Horvath, L. (2001). Change-point detection in angular data. *Annals of the Institute of Statistical Mathematics*, 53(3), 552-556.

Jammalamadaka, S. R., & SenGupta, A. (2001). *Topics in circular statistics*. Singapore: World scientific.

Jiang, Q. (2009). *On fitting a mixture of two von-Mises distributions, with applications*. M.Sc Project, Department of Statistics and Actuarial Science, Simon Fraser University, Canada.

Kato, S., & Jones, M. C. (2010). A family of distributions on the circle with links to, and applications arising from mobius transformation. *Journal of American Statistical Association*, 105 (489), 249-262.

Kong, L. B., Lekawa,M., Navarro, R. A,. McGrath, J., Cohen, M., Margulies, D. R., &Hiatt, J. R. (1996). Pedestrian-motor vehicle trauma: an analysis of in-

jury profiles by age. *Journal of the American College of Surgeons*, 182, 17-23.

Lombard, F. (1986). The change-point problem for angular data: A nonparametric approach. *Technometrics*, 28, 391-397.

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics*. Chichester,Wiley.

Mooney, J. A., Helms, P. J., & Jollife, I. T. (2003). Fitting mixtures of von-Mises distributions: A case study involving sudden infant death syndrome. *Computational Statistics & Data Analysis*, 41, 505-513.

Mullen, K. M., Ardia, D., Gil,D. L., Windover, D., & Cline, J. (2011) . DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software*, 40(6), 1-26.

Owens, D. A., & Sivak , M. (1993). *The role of reduced visibilityin night time road fatalities*. Report no. UMTRI-93-33, University of Michigan. Transportation Research Institute, U.S.A.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42, 523-527.

Plainis, S, Murray, I. J., & Pallikaris, I. G. (2006). Injury prevention, 12, 125-128. doi: 10.1136/ip.2005.011056.

Roy, A., Parui, S. K., & Roy, U. (2012). A mixture model of circular-linear distributions for color image segmentation. *International Journal of Computer Applications* (0975 - 8887), 58(9), 6-11.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.

Sengupta, A., & Laha, A. K. (2008a). A likelihood integrated method for exploratory graphical analysis of change point problem with directional data. *Communications in statistics. Theory and methods*, 37(11-12), 1783-1791.

Sengupta, A., & Laha, A. K. (2008b). A Bayesian analysis of the change-point problem for directional data. *Journal of Applied Statistics*, 35(6), 693-700.

von Mises, R. (1918). UJber die 'Ganzzahligheit' der Atomgewichte und vermandteFragen. Phys. Z. 19, 490-500.

Zhang, H., & Huang, Y. (2015). Finite mixture models and their applications: A review. *Austin Biometrics and Biostatistics*, 2(1), 1013, 1-6.

Article can be accessed online at http://www.publishingindia.com

Text Analytics Framework using Apache Spark and Combination of Lexical and Machine Learning Techniques

Anuja Prakash Jain*, Padma Dandannavar**

Abstract

Today, we live in a 'data age'. The sudden increase in the amount of user-generated data on social media platforms like Twitter, has led to new opportunities and challenges for companies that strive hard to keep an eye on customer reviews and opinions about their products. Twitter is a huge fast emergent micro-blogging social networking platform for users to express their views about politics, products sports etc. These views are useful for businesses, government and individuals. Hence, tweets are used in this framework for mining public's opinion. Sentiment analysis is a process of naturally recognizing whether a user-generated content expresses positive, negative or neutral opinion about an entity (i.e. product, people, topic, event etc). The traditional analytics tools are costly and are not built to handle Big data. Hadoop, though being a popular framework for data intensive applications, does not perform well on iterative process (like data analysis) due to the cost paid for data reloading from disk for each iteration. This paper proposes a Text analysis framework for twitter data using Apache spark and hence is more flexible, fast and scalable. The proposed framework is also domain independent as it uses a hybrid approach by combining supervised machine learning algorithms (Naïve Bayes and decision tree machine learning algorithms) and lexicon approach (pattern analyzer) for sentiment classification thereby comparing various supervised learning models and using the one with highest accuracy for predicting sentiment.

Keyword: Sentiment Analysis, Machine Learning, Lexical Approach, Apache Spark, Natural Language Processing, Twitter

Introduction

Twitter is a huge, fast, emergent, popular, micro-blogging social networking platform for users to express their views about politics, products sports etc. Here, clients send messages (a.k.a., tweets) to a system of contacts from a wide assortment of gadgets or sites. A tweet is a content predicated post and has just 140 characters, which is around the length of a typical newspaper headline or subhead (Milestein, 2008). Twitter is a "what's-happening-right-now" social network and hence tweets are valuable sources for businesses, government and individuals to determine public's opinion or sentiment about an entity (product, people, topic, event etc). But, the volume of tweets produced by Twitter everyday is very vast (21 million tweets per hour, as measured in 2015). Hence there is a need to automate the process of sentiment analysis so as to ease the tasks of determining public's opinions without having to read millions of tweets manually. This process of analyzing and summarizing user's views on a particular entity is usually called Sentiment Analysis or Opinion Mining which is an extremely fascinating and prominent space for analysts these days.

Text analysis includes data retrieval, lexical analysis to study word recurrence appropriations, pattern recognition, labeling/annotation, data extraction, data mining techniques, visualization, and predictive analytics. The general objective is, basically, to transform content into information for investigation, by means of use of Natural language processing (NLP) and analytical methods. Sentiment analysis is a process of automatically identifying whether a user-generated content expresses

* M.Tech Student, Computer Science and Engineering, Visvesvaraya Technological University, Karnataka, India. Email: jainanu04@gmail.com

** Assistant Professor, Padma Dandannavar, Computer Science and Engineering, Gogte Institute of Technology, Belgaum, Karnataka, India. Email: padmad@git.edu

positive, negative or neutral opinion about an entity (i.e. product, people, topic, event etc). Sentiment classification can be done at Document level, Sentence level and Aspect or Feature level (Vohra, 2013). In Document level the whole document is used as a basic information unit to classify it either into positive or negative class. Sentence level sentiment classification classifies each sentence first as subjective or objective and then classifies it into positive, negative or neutral class. There is not much difference between the above two methods as sentence is just a short document. Aspect or Feature level sentiment classification deals with identifying and extracting product features from the source data (Vohra, 2013).

There are few methodologies for sentiment analysis: *Machine learning based approach* (ML) uses several machine learning algorithms (supervised or unsupervised algorithms) to classify data (Neethu, 2013). Here, two datasets are needed: training and a test dataset. A supervised learning classifier utilizes the training set to learn and train itself w.r.t the differentiating characteristics of text, and a test set is utilized to check the performance of the classifier. *Lexicon based approach* uses a dictionary containing positive and negative words to determine the sentiment polarity. It deals with counting the number of positive and negative words in the text. If the text consists of more positive words, the text is assigned a positive score. If there is more number of negative words then the text is assigned a negative score. If the text contains equal number of positive and negative words then it is assigned a neutral score. To determine whether a word is positive or negative, an opinion lexicon (positive and negative opinion words) is built. *Hybrid based approach* uses a combination of both ML and lexicon based approach for classification. The drawback of Machine learning based approach is that it needs a huge training data which is very difficult to obtain. Also, manually labeling the data is a very tedious job. Lexicon based approach has a disadvantage that the strength of the sentiment classification depends on the size of the lexicon (dictionary). As the size of the lexicon increases this approach becomes more erroneous and time consuming. As mentioned in (Zang, 2011), lexicon based approach has high precision and low recall. Hence combining it with a machine learning classifier can improve the recall and accuracy of the algorithm. This paper focuses on the proposition of combining the two approaches into a hybrid model in order to mitigate the drawback of these two approaches by using the lexicon-based classifier for the task of annotating the training data for the learning-based classifier, this technique leverages the learning-based classifier's performance while taking advantage of the lexicon-based classifier's effortless setup process (Zang, 2011).

The amount of user generated data available online is astronomically immense. Considering the span of information starting 2011, i.e. 30 million, we can anticipate that it will increase to over billions by 2017. Consequently we have to guarantee that the framework is adaptable and *fast* to support any measure of information. This project develops a framework that is both scalable and fast. This is accomplished by using Apache Spark. Apache Spark is a quick and universally useful cluster computing system. Spark runs programs up to 100x speedier than Hadoop MR in memory, or 10x quicker on disk. Machine Learning Library (mllib) is available in Apache Spark (Bhuvan, 2015). With the help of this, various classification models are built and the model with the highest accuracy is used to predict the user sentiment on twitter data.

The Proposed Technique

This section describes the proposed framework. The framework uses hybrid approach for analysis. Figure 1 gives an architectural overview of the proposed technique for twitter sentiment analysis.

A. Data Collection

Twitter allows researchers to collect tweets by using a Twitter API. One must first create a twitter application to obtain twitter credentials (i.e. API key, API secret, Access token and Access token secret) which can be obtained from twitter developer site. These credentials should be kept private as they provide application access to twitter on behalf of your account. The twitter application then receives an "OAuth access token" which ensures that authorized calls are made to Twitter API's. OAuth is a method for permitting users to approve third party applications to access their account data without expecting to share delicate data like password. Then, install a twitter library to connect to the Twitter API. We used Twitter Search API to extract tweets based on a search query (e.g. #Apple , "Apple Products") specified by the user. It works similar to the search function provided by twitter web client or twitter mobile. We used tweepy twitter library to extract tweets. Twitter API has certain rate limits on how many requests an application can make to any API resource within given time window. Twitter provides 15 minutes window and allows the user to access tweets of past 7 to 8 days if the type of result is specified as 'recent'.

The twitter data is obtained in Json form containing tweets and information about the users.

The important fields are explained below:

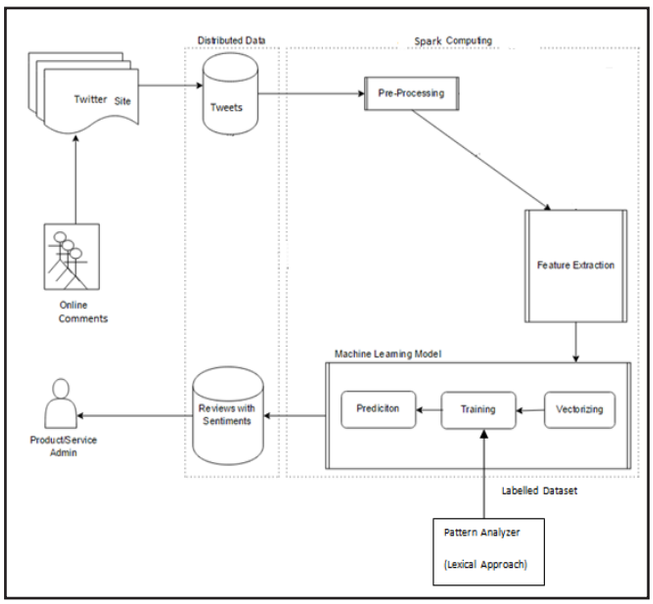


Fig 1: Architecture of Twitter Sentiment Analysis using Apache Spark and Combination of Lexical and Machine Learning Technique

1. Created at: Time at which the user posted the tweet.
2. Entities: Several fields like URL, Hashtags, user mentions parsed from text.
3. Id: A unique identifier for this tweet.
4. Retweet count: Indicates the number of times this tweet is retweeted.
5. Source: indicates the source of a tweet (device or website).
6. Text: shows the actual tweet as posted by the user containing external web links (e.g. http://amze.ly/8K4n0t) (Zang, 2011), hashtags (e.g. #Apple , used to filter tweets based on a topic), user names (e.g. @user1, indicates that the tweet is a reply to a user named user1) and the user comment.
7. User: containing information about the user like user id, user's profile image, description of the user etc.

Data Pre-processing

Before we perform the sentiment analysis on twitter data the data should be brought into proper form and sentiment relevant features need to be extracted. The text field of twitter data, as described above contains external links, URLs and tweets. Processing a natural language is very difficult and trying to determine the user sentiment is even more difficult as users make sarcastic comments. The data preprocessing step includes the following:

1. Removal of external links: URLs and user names are not useful for determining the sentiment analysis. Eg. Removing "http://act.org" and "@valaafshar" in figure 2
2. Removal of duplicate tweets: Twitter data may contain redundant tweets and retweets which need to be removed.
3. Spelling Correction: Social media tweets may contain incorrect spellings. Spelling of the erroneous words can be rectified and predicated on automated selection of more probable word.
4. Case Conversion: All words are transformed into lower case in order to eradicate the distinction between "Text" and "text" for further processing.
5. Stop-words Removal: The commonly used words like a, an, the, has, have etc which carry no meaning i.e. do not help in determining the sentiment of text while analyzing should be removed from the input text. E.g Removing words like "it","have","and" etc as shown in figure 2.

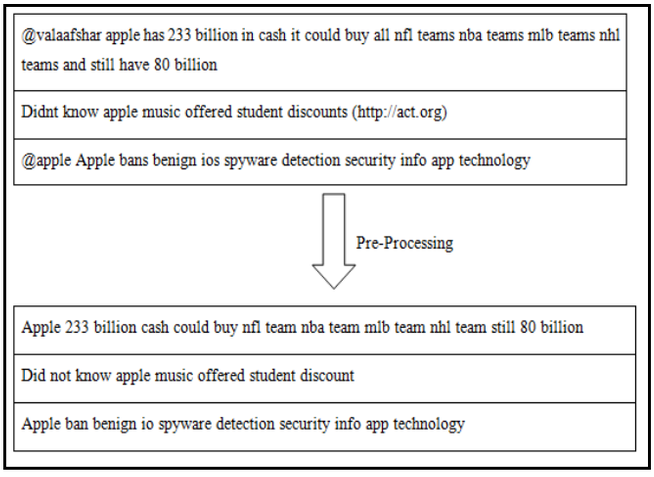


Fig 2: Data pre-processing example

6. Punctuation Removal: Punctuation marks such as comma or colon often carry no meaning for the textual analysis hence they can be removed from input text.

7. Stemming: Stemming customarily refers to a simple process that chops off the terminuses of words to abstract derivational affixes. E.g converting "teams" to "team".
8. Lemmatization: Deals with abstraction of inflectional terminuses only and to return the base or dictionary form of a word, which is termed as the lemma. E.g. Converting words like "are" "am" "is" to "be".

The figure below shows sample tweets and the results obtained after pre-processing.

The proposed method makes use of NLTK (Natural Language Tool Kit) to carry out pre-processing.

Feature Extraction and Vectorizing

Once the tweets are pre-processed, we need to extract features relevant for sentiment analysis. This framework uses Hashing TF-IDF algorithm to extract features. TF-IDF (*term frequency-inverse document frequency)* is frequently utilized in information retrieval and text mining. This TF-IDF weight is a statistical measure used to assess the importance of a word in a document (tweet). The Hashing TF-IDF algorithm implemented in the proposed framework using Apache Spark is:

Algorithm: Hashing TF-IDF (tweet)
Input: RDD of tweet
Output: TF-IDF weight of each term in the tweet
Hash (t) – hash function for term t
num_of_terms – total number of terms in a tweet
Freq (*t* ,*tweet*) – Calculates number of times t appears in a *tweet*
d(*t*)- Number of tweets that contain term *t*
m-Total number of tweets
Vector.sparse (arg) - Returns a Sparse Vector of arg

1. For each term *t* in *tweet*:
    a. Calculate index of *t*

        index = hash(*t*) % No. of terms

    b. Frequency[index] = Freq (*t,tweet*)
2. Vector = Vector.Sparse(No. of terms, Frequency[ ])
3. Data = RDD of Vectors
4. Calculate IDF for each vector in Data using the formula:

        idf = log((m + 1) / (d(*t*) + 1))

5. Calculate TF-IDF weight using:

    TF-IDF = Frequency[*t*]*idf

This approach eschews the need to compute a global term-to-index map, which can be extravagant for a sizably voluminous corpus, but it suffers from hash collisions, where different raw features may become identically tantamount term after hashing. In order to avoid this, the number of buckets in the hash table is increased to 1,048,576. A sample hashed feature vector is as shown below:

SparseVector (50000, {20420: 0.8755, 23406: 0.9445, 27595: 3.989, 39565: 3.989})

Here, feature vector is a Sparse Vector and term '50,000' denotes the bucket size or feature dimension. The terms of the form '20420: 0.8755' denotes 'hash-value: feature'

Building A Training Dataset.

The lexicon-based approach is used to build the training data. The training data consists of tweets labeled by lexical-based pattern analyzer (Zang, 2011). Using training data provided by lexicon-based method has following advantages:

- Mitigating the labor-intensive and time consuming process of manually annotating training data.
- Lexicon based approach have low recall. By using hybrid approach we can achieve higher accuracy and recall. (Zang, 2011) (Fredrick, 2015).
- The proposed framework is domain independent as the classifier is trained on the fly based on the output of lexical-based approach and not on a domain-specific manually labeled dataset.

Machine learning algorithms require huge training data for better performance. Suppose, on an average if a human takes around 10 seconds to classify one tweet, then he would take 15000 seconds (4 hours) to classify 1500 tweets. Though a machine learning approach performs slightly better than the proposed hybrid approach, the difference in performance might no longer be worth the inconvenience of acquiring training data, making the hybrid model an appealing alternative with a more beneficial trade-off between performance and convenience. (Fredrick, 2015). Nowadays, large organizations are desperately in need of fast approximate results rather than accurate results

to take important decisions faster. Hence, the proposed framework fulfils this need by making use of hybrid approach and Apache Spark framework.

Predicting User Sentiment

Machine learning classifiers are trained on the training data obtained from lexicon approach and then they are tested on test dataset. The machine learning model classifies the tweets as positive, negative and neutral. Their performances are compared based on various performance measures like accuracy, precision, recall and F1-score. The machine learning algorithms used in this framework are explained in the next section.

Machine Learning Algorithms for Sentiment Classification

Multinomial Naïve Bayes

The Naive Bayes classifier is the easiest of all (as the name proposes) and extremely viable for text classification as it figures the posterior probability of a class, in view of the dispersion of the words (features) in the document. We use multinomial naive bayes as we classify a tweet into 3 different classes (positive, negative, and neutral). Here, each feature is denoted as a term whose value is the frequency of the term. The Multinomial naive bayes algorithm is as shown below:

```
TrainMultinomialNB(ℂ, 𝔻)
 1  V ← ExtractVocabulary(𝔻)
 2  N ← CountDocs(𝔻)
 3  for each c ∈ ℂ
 4  do N_c ← CountDocsInClass(𝔻, c)
 5     prior[c] ← N_c / N
 6     text_c ← ConcatenateTextOfAllDocsInClass(𝔻, c)
 7     for each t ∈ V
 8     do T_ct ← CountTokensOfTerm(text_c, t)
 9     for each t ∈ V
10     do condprob[t][c] ← (T_ct + 1) / (Σ_t' (T_ct' + 1))
11  return V, prior, condprob

ApplyMultinomialNB(ℂ, V, prior, condprob, d)
 1  W ← ExtractTokensFromDoc(V, d)
 2  for each c ∈ ℂ
 3  do score[c] ← log prior[c]
 4     for each t ∈ W
 5     do score[c] += log condprob[t][c]
 6  return arg max_{c∈ℂ} score[c]
```

Multinomial naive bayes perform better when trained on huge dataset.

Decision Tree

Decision trees in Spark MLlib are greedy algorithms and scale gracefully to distributed setting. A decision tree model is trained using training dataset and model builds a top-down tree (hierarchical if-else statements) which can then be used to predict unseen data. The decision tree performs binary partition of the feature space recursively. The tree avariciously picks every segment by selecting the best split from an arrangement of conceivable parts, with a specific end goal to expand the data pick up at a node of tree. We use *Gini Impurity* to measure the homogeneity of labels at each node.

These two algorithms are tested on test data and their performances are compared based on accuracy, precision, recall, F1-score.

Advantages of Proposed System

- Mitigating drawbacks of Hadoop Map Reduce: Hadoop does not perform well on real-time iterative processes (like data analysis) due to the cost paid for data reloading from disk for each iteration. The proposed framework uses in memory Apache Spark for overcoming these drawbacks.
- Overcoming tedious job of manually labeling huge training data: The framework uses a pattern analyzer for labeling the training data.
- Domain Independence: Since the framework uses a hybrid approach (i.e. lexicon approach to label the training data and machine learning approach to predict sentiments) , it can perform analysis on any type of dataset independent of any specific domain(like telecom, Banking)

Limitation: The framework relies on pattern analyzer for building the training data. The accuracy of pattern analyzer may not be as good as that of manually labeled dataset. However, there is a tradeoff between ease of obtaining labeled training data and accuracy. Analysts believe that in today's fast-paced world, obtaining "approximate" results fast is better than obtaining "accurate" results late. Hence based on this belief, though the framework may slightly lack in accuracy, it is very apt for performing real-time analysis thereby aiding the organizations to take important decisions quickly.

Results

We used a dataset containing 4000 tweets for search query "Apple Products". 70% of the dataset was used for training while 30% was utilized for testing the performance of Naïve Bayes and Decision tree algorithms. The performance measured in terms of precision, recall, F1Score, Accuracy is shown in the figure 1. The figure shows that Decision tree algorithm has an accuracy of 100 percent and Naïve Bayes has an accuracy of 86 percent. The figure also shows precision, recall and F1Score.

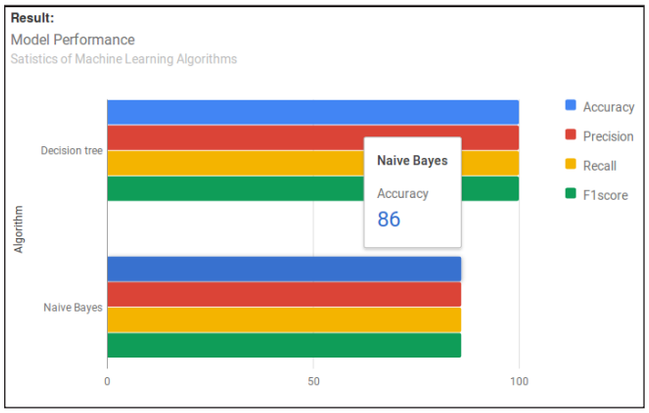


Fig. 1: Performance of Naïve Bayes and Decision Tree

The result of sentiment analysis is shown in the form of a pie chart. Figure 2 shows that about approx 20% tweets are negative for the search query "Apple Products".

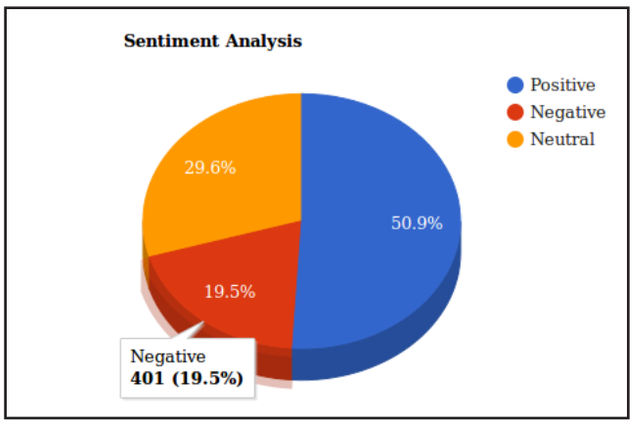


Fig. 2: Sentiment Analysis using Number of Positive, Negative and Neutral Tweets for Queries "Apple Products"

The framework also shows the sources of tweets which are obtained from the 'source' key in twitter Json.

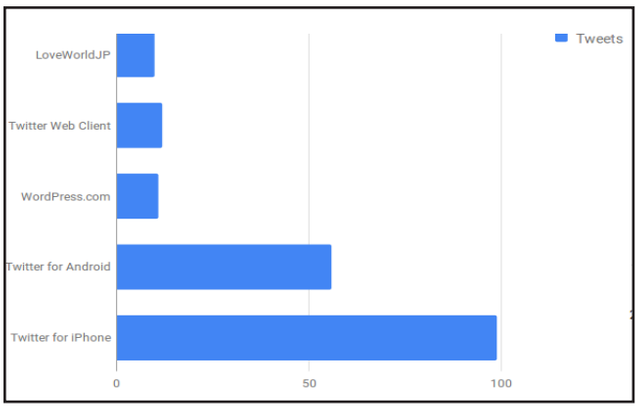


Fig. 3: Sources of Tweets

The country-wise location of tweets is shown using the world graph below. The graph shows that maximum people who are tweeting about Apple Products are from USA.

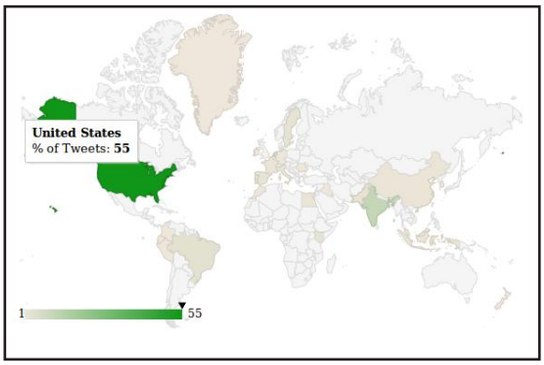


Fig. 4: Country-wise Location of Tweets

It is also useful to know on what topic are people talking about. The framework gives an idea of the topic of discussion using a word cloud. The word cloud shows that most of the people are talking about "apple" while few others are talking about mac and ios.

Fig. 5: Word Cloud of Important Keywords

Conclusion

Sentiment analysis can be performed using lexicon based approach, machine learning based approach or hybrid approach. The lexicon-based approach is used to build the training data for mitigating the labor-intensive and time consuming process of manually annotating training data. The proposed framework is domain independent. The framework performs sentiment analysis using Naive Bayes and Decision tree algorithms. The results show that Decision tree performs extremely well showing 100% accuracy, precision, recall and F1Score. The framework also shows the sources and location of tweets along with the important keywords depicting the topic of discussion. The proposed text analytics framework is also real-time, fast, scalable, and reliable as we use Apache Spark framework.

References

Hassan, A., & Medhat, W. (2014). Sentiment analysis algorithms and applications: A survey. *Shams Engineering Journal*, 5, 1093-1113.

Andrea, A. D., & Ferri, F. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3).

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.* Hewlett-Packard Development Company.

Vohra, S. M., & Teraiya, J. B. (2013). A comparative study of sentiment analysis techniques. *Journal of Information, Knowledge and Research in Computer Engineering,* 2(2), 313-317.

Zhang, M.-L., Pe˜na, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. *Elsevier, Information Science Journal.*

Neethu, M. S., & Rajashree, R. (2013). *Sentiment Analysis in Twitter using Machine Learning Techniques*. Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT).

Turney, P. D. (2002). *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews.* In proceedings of 4th annual meetings for computational linguistics, 417-424.

Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B., & Magoulas, R. (2008). *Twitter and the micro-messaging revolution: Communication, connections*. An OReilly Radar Report, p. 54.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 79& 86.

Ko, E. H., & Klabjan, D. (2014). *Semantic Properties of Customer Sentiment in Tweets*, 28th International Conference on Advanced Information Networking and Applications Workshops.

Liu, B. (2012). *Sentiment analysis and opinion mining* (18-19, 27-28, 44-45, 47, and 90-101). Morgan and Claypool Publishers.

Cho, S. H., & Kang, H.-B. (2012). *Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary*. IEEE International Conference on Conference on consumer Electronics (ICCE), 717-718, 2012.

Kanakaraj, M., & Guddeti, R. R. (2015). *NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers.* 3rd International Conference on Signal Processing, Communication and Networking.

Mane, S. B., Sawant, Y., Kazi, S., & Shinde, V. (2014). Real time sentiment analysis of twitter data using Hadoop. *International Journal of Computer Science and Information Technologies,* 5(3), 3098-3100.

Rajurkar, G. D., & Goudar, R. M. (2015). A speedy data uploading approach for Twitter Trend And Sentiment Analysis using Hadoop. International Conference on Computing Communication Control and Automation.

Bhuvan, M. S., & Rao, V. D. (2015). *Semantic Sentiment Analysis Using Context Specific Grammar.* International Conference on Computing, Communication and Automation (ICCCA2015).

Article can be accessed online at http://www.publishingindia.com

The Impact of the Scale Elements Alteration on Priorities in Analytic Hierarchy Process Technique

Mohammad Azadfallah*

Abstract

The present study, presents a comparative analysis of different measurement scales adopted in Analytic Hierarchy Process (AHP), by testing them versus a problem with a known composite answers. Then experimentally, the impact of the different measurement scale elements alteration from three aspects: 1. The limited scale upper bound (up to 9), 2. Changing the scale parameters (a parameters), and 3. Changing the system numbers (from 1, 3…9; to 2, 4…10) on priorities are investigated. The results show that the linear measurement scale has the best performance in comparison to other scales.

Keyword: AHP, Measurement Scale, Scale Elements Alteration

Introduction

Analytic Hierarchy Process (AHP) has been a tool in the hands of decision makers and researchers since its invention; it is still the most widely used multi-criteria decision making method (Turskis et al., 2009). AHP is based on three basic principles: decomposition, comparative judgments, and hierarchic composition or synthesis of priorities. The decomposition principle is applied to structure a complex problem into a hierarchy of clusters, sub-clusters, sub-sub-clusters and so on. The principle of comparative judgments is applied to construct pair wise comparisons of all combinations of elements in a cluster with respect to the parent of the cluster. These pair wise comparisons are used to derive 'local' priorities of the elements in a cluster with respect to their parent. The principle of hierarchic composition or synthesis is applied to multiply the local priorities of elements in a cluster by the 'global' priority of the parent element, producing global priorities throughout the hierarchy and then adding the global priorities for the lowest level elements (the alternatives), (Forman and Selly, 2001). One of AHP's strengths is the possibility to evaluate quantitative as well as qualitative criteria and alternatives on the same preference scale of nine levels. These can be numerical, verbal, or graphical (Ishizaka and Labib, 2009). Theoretically there is no reason to get restricted only to these numbers. Therefore, other scales have been proposed (Ishizaka et al., 2011). Since, the main goal of the present research is to evaluate the different measurement scales and scale elements alteration on priorities in AHP techniques.

The paper is organized as follows. In section 2; the AHP, section 3; measurement scale and section 4; literature is reviewed. Numerical example is provided in section 5; the paper is concluded in section 6.

Analytic Hierarchy Process (AHP)

AHP is an intuitive method for formulating and analyzing decisions. AHP has been applied to numerous practical problems in the last few decades. Because of its intuitive appeal and flexibility, many corporations and governments routinely use AHP for making major policy decisions (Ramanathan, 2001). It is not the purpose of this paper to explain in detail the AHP methodology. See for instance Saaty (2000). A brief discussion of AHP is provided in this section.

The AHP method uses the pair wise comparisons and eigenvector methods to determine the a_{ij} values and also the criteria weights W_j. In this method; a_{ij} represents the relative value of alternative A_i when it is considered in terms

* Researcher, Business Studies and Development Office, Saipayadak (Saipa after sales services organization), Islamic Republic of Iran. Email: m.azadfallah@yahoo.com

of criterion C_j. In the original AHP method, the a_{ij} values of the decision matrix need to be normalized vertically. That is, the elements of each column in the decision matrix add up to 1. In this way, values with various units of measurement can be transformed into dimensionless ones. If all the criteria are benefit criteria (that is, the higher the score the better the performance is), then according to the original AHP method, the best alternative is the one that satisfies the following expression:

$$P^*_{AHP}=\max_i P_i=\max_i \sum_{j=1}^{n} a_{ij}W_j, \text{ for } i=1, 2, 3\ldots m.$$

From the above formula, it can be seen that the original AHP method uses an additive expression to determine the final priorities of the alternatives in terms of all the criteria simultaneously (Wang, 2007). Generally, the purpose of the AHP is to assist people in organizing their thoughts and judgments to make more effective decisions (Saaty, 2000).

Measurement Scale

Almost all sciences use numbers. These numbers appear throughout all levels of the complex chain of mathematical, logical, and heuristic analyses that constitute scientific explanation and argumentation. Usually the first place they appear is in the quantification of empirical concepts. This step is usually called measurement (Narens, 1981). Measurement is any set of rules for assigning numbers that are attributed to objects (Saaty, 2004). The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurements (Stevens, 1946). The only rule not allowed would be random assignment, for randomness in effect amounts to a non-rule (Luce, 1997). In his 1946 and 1951 publications Stevens singled out four groups of transformations on the real or positive real numbers as relevant to measurement: one-to-one, strictly monotonic increasing, affine, and similarity, and he introduced the corresponding terms of nominal, ordinal, interval, and ratio to refer to the families of homomorphism, or scales, related to these groups (Narens and Luce, 1986).

A commonly used measurement scales in the AHP is the ratio scale (Vachajitpan, 2004). Perhaps the most significant aspect of the AHP is in its use of ratio scales (Saaty, 2000). The measurement scale proposed by Saaty in the AHP is a 1 to 9 point scale. It is used to indicate the number of times one criterion is better than other criteria in the pair wise comparison. The reverse relationship is represented by an inverse of the assigned value. Thus, it is impossible to have a zero or a negative value in the AHP scale (Vachajitpan, 2004).

In a judgment matrix, instead of assigning two numbers W_i and W_j (numbers that generally we do not know), as one does with tangibles, and forming the ratio W_i / W_j we assign a single number drawn from the fundamental scale of absolute numbers shown in table 1, to represent the ratio $(W_i / W_j)/1$, (Saaty, 2005). Theoretically there is no reason to be restricting to these numbers. Therefore, other scales have been proposed (table 2), (Ishizaka et al., 2011).

Table 1: The fundamental scale of absolute numbers

Intensity of Importance	*Definition*	*Explanation*
1	Equal importance	Two activities contribute equally to the objective.
2 3	Weak Moderate importance	Experience and judgment Slightly favor one activity over another.
4 5	Moderate plus Strong importance	Experience and judgment Strongly favor one activity over another.
6 7	Strong plus Very strong or demonstrated importance	An activity is favored very strongly over another; its dominance demonstrated in practice.
8 9	Very, very strong Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation.
Reciprocals of above	If activity I has one of the above non zero numbers assigned to it when compared with activity j, then j has the reciprocal value when compared with i	A reasonable assumption.
Rationales	Ratio arising from the scale	If consistency were to be forced by obtaining n numerical values to span the matrix.

Ref. Saaty (2005), p. 356.

In general, evaluating of the impact of the different measurement scale elements alteration on priorities in AHP is the aim of this paper.

Table 2: Different Measurement Scales

scale	*definition*	*parameters*
linear	C=a.x	a>0;x=1,2,…,9
power	$C=x^a$	a>1;x=1,2,…,9
geometric	$C=a^{x-1}$	a>1;x=1,2,…,9
logarithmic	$C=\log_a^{(x+1)}$	a>1;x=1,2,…,9
Root square	$C=\sqrt[a]{x}$	a>1;x=1,2,…,9
Inverse linear	C=9/(10-x)	a>1;x=1,2,…,9
balanced	C=w/(1-w)	W=0.5,0.55,0.6,…,0.9

Ref. Ishizaka et al., (2011), p. 4.

Literature Review

In the current literature, there are several examples where different measurement scale is used and compared in the choice of the final solution. Here, we will mention some of them. Poyhonen et al., (1997) performed a comparative study in which subjects were requested to quantify verbal ratio statements by adjusting the heights of visually displayed bars. Salo and Hämäläinen (1997) applied multi attribute value theory as a framework for examining the use of pair wise comparisons in the AHP. Next, it is demonstrated that the AHP can be modified so as to produce results similar to those of multi attribute value measurement. Then, the new balanced scales to improve the sensitivity of the AHP ratio scales are proposed. Triantaphyllou et al., (1998) provided a comprehensive survey of some methods for eliciting data (measurement scale) for MCDM (Multiple Criteria Decision Making) problems and also for processing such data. Sato (2001) studied to find the scale (i.e. linear and power scale) of the AHP appropriate for representing decision maker's perception. Result indicated that the power scale is preferable to the linear scale as the judgment scale. In Shinohara et al., (2001) various methods, such as the eigenvector method, the geometric mean method, and the entropy method have been proposed to estimate a weight vector from pair wise comparison data. The results indicated that, when a decision maker decides each element of a pair wise comparison matrix on the basis of linear scale, the entropy method is expected to produce a weight vector that is closest to the true weight vector. On the contrary when a decision maker decision is on the basis of exponential scale, the eigenvector method and the geometric mean method are expected to produce weight vectors closer to the true weight vector. Vachajitpan (2004) developed a different model based on the least square principle to apply to situations where either the ratio or the intervals scaling method are used. Wedley (2007) studied the role of natural zero in scale for generating priorities. Monat (2009) proposed the use of global scales instead of local scales. (Because, using the local scales tends to overemphasize the small differences in attribute measures). Cox (2009) used a graph for interpreting of multidimensional data. So, at first the much dissimilarity generated in the ANP (Analytic Network Process) is analyzed using individual differences scaling. Secondly the single sets of dissimilarities that arise from the AHP are analyzed using multidimensional scaling. Ishizaka et al., (2011) demonstrated that the aggregation method of local priorities and the measurement scale in AHP has a strong influence on the selection of the compromise and therefore on the degree of concordance with the utility theory. Munshi (2014) proposed a method by which likert scales may be tailored for any given instrument and semantics. The method consists of performing a pretest using unmarked lines as scales, measuring the distances marked by the respondents, and using cluster analysis to determine the best placement of scale points for the actual study. However, as far as we know, a few experimental studies have addressed a fundamental problem discussed in this paper (altering the measurement scale elements). I.e. Triantaphyllou et al., (1994) used two evaluative criteria: 1. the ranking yielded when the CDP (the Closest Discrete Pair wise) matrix is used should not demonstrate any ranking inversions when the CDP ranking is compared with the ranking derived from the RCP (the Real Continuous Pair wise) matrix. 2. The ranking yielded when the CDP (the Closest Discrete Pair wise) matrix is used should not demonstrate any ranking indiscrimination when the CDP ranking is compared with the ranking derived from the RCP (the Real Continuous Pair wise) matrix, to examine a total of 78 scales which can be derived from two widely used scales (altering the two measurement scale parameters): 1. original scale (linear scale proposed by Saaty) and 2. Exponential scale. Results demonstrated that there is no single scale that can always be classified as the best or the worst scale for all cases. Ji and Jiang (2003) first reviewed and compared different scales from different aspects. Then discussed the transitivity of AHP scales and derived a scale based on the transitivity. Next, proposed two approaches for determining the parameter of the derived transitive scale. The result indicated that, proposed scale is quite simple and practicable. This paper proposes a new approach as discussed below.

Numerical Example

To derive priorities, the verbal comparisons must be converted into numerical ones (Ishizaka and Labib, 2009).

A comparison of the different measurement scales (based on the formula in table 2) is given in table 3.

Table 3: Different Scales for Comparing Two Alternatives

Scale type	*Values*								
linear	1	2	3	4	5	6	7	8	9
power	1	4	9	16	25	36	49	64	81
geometric	1	2	4	8	16	32	64	128	256
Logarithmic	1	1.58	2	2.32	2.58	2.81	3	3.17	3.32
Root square	1	1.41	1.73	2	2.23	2.45	2.65	2.83	3
Asymptotical*	1	0.12	0.24	0.36	0.46	0.55	0.63	0.70	0.76
Inverse linear	1	1.13	1.29	1.5	1.8	2.25	3	4.5	9
Balanced	1	1.22	1.5	1.86	2.33	3	4	5.67	9

Ref. Ishizaka and Labib (2009), p. 209.

*. Accordance Vachajitpan (2004), "it is impossible to have a zero or negative value in the AHP scale". Therefore, the asymptotical scale will not be discussed here.

In this section, we study how by altering the different measurement scale elements from the different aspects we can analyze the existing measurement scales. Experimentally, three tests include: 1. the limited scale upper bound (up to 9), 2. Changing the scale parameters (a parameters), and 3. Changing the system numbers (from 1, 3…9; to 2, 4…10) provided. Next, via numerical example the impact of different measurement scale on priorities and their performance in terms of each one of the test criterion are investigated. Here, two points are noteworthy. First, in accordance to Wedley (2001, p. 551):

> "At the 5th international symposium of the Analytic Hierarchy Process in Kobe Japan, Thomas Saaty suggested to the author that the controversy regarding correct synthesis modes for the AHP should be tested with problems with known true values …".

Second, in accordance to Saaty (2000, p. 455) have:

> "In the AHP one needs to be careful with criteria measured on the Same absolute scale. Criteria measured in dollars are a common Example of this. The priority of each criterion must be equal to The sum of the measurements of its alternatives divided by the Sum of the measurements of the alternatives with respect to all These criteria. Only then can one normalize the measurements Of the alternatives, weight them by these priorities and add to Obtain the relative weights of the alternatives with respect to All these criteria".

Since, the use of problems with known answers is the aim of this paper. To illustrate these basic ideas, and assuming that all of the criteria are expressed in the same unit, a simple 4.4 decision matrix is presented (table 4).

Table 4: Problem with Known Weights

Cri. / Alt.	*C1*	*C2*	*C3*	*C4*	*total*	*True weights*
A1	1	5	9	3	18	0.225
A2	7	3	1	9	20	0.250
A3	3	3	9	7	22	0.275
A4	5	7	3	5	20	0.250
total	16	18	22	24	80	-

In the absence of any other standards, the solution provided by this approach (A3, .275 > A2, .250 = A4, .250 > A1, .225), was used as the standard. Since, this could cause some bias in the final result.

Table 5: AHP Results for Linear Scale

Cri. / Alt.	*C1 16/80 =.200*	*C2 18/80 =.225*	*C3 22/80 =.275*	*C4 24/80 =.300*	*Composite priorities*
A1	0.063	0.278	0.409	0.125	0.225
A2	0.438	0.167	0.045	0.375	0.250
A3	0.188	0.167	0.409	0.292	0.275
A4	0.313	0.389	0.136	0.208	0.250

A comparison of the test results is given in table 6. i.e. for linear measurement scale (based on table 4):

Notes: According to axiom 3 of AHP, the criteria are assumed to be independent of the alternatives (Wedley, 2001). Here, we violated this property.

Table 6: AHP Results for Different Measurement Scales

Scale type	*Priorities (rank and intensity)*
linear	**A3 > A2 = A4 > A1** .275 .250 .250 .225
power	A3 > A2 > A1 > A4 .289 .273 .227 .211
geometric	A3 > A2 > A1 > A4 .318 .316 .269 .097
logarithmic	A3 > A4 > A2 > A1 .267 .263 .241 .230
Root square	A4 > A3 > A2 > A1 .263 .262 .241 .233
Inverse linear	A3 > A2 > A1 > A4 .292 .287 .263 .158
balanced	A3 > A2 > A1 > A4 .288 .279 .249 .183

Findings

1. Different measurement scales for deriving priorities, can lead to different results.
2. As seen from the table, the linear measurement scale results are the same (rank and intensity) as displayed in the last column of table 4 (standard).
3. It is remarkable to observe that in this illustrative example using the all of the scales, except Root square, the A3 ranking is best.

Test Criterion

The Limited Scale Upper Bound (up to 9)

In this section, the all of the measurement scales are upper bounds, by altering scale parameters restricted to the 9 (table 7). Because, the base (upper bound) for original AHP scales is nine. Then, the behaviors of these methods (scales) by their impacts on priorities are investigated.

Table 7: Modified Measurement Scales

Scale type	*Scale parameter*	*values*								
Linear*	-	1	2	3	4	5	6	7	8	**9**
power	a=1.0001	1	2	3	4	5	6	7	8	**9**
geometric	a=1.3161	1	1.32	1.73	2.28	3	3.95	5.20	6.84	9
Logarithmic	a=1.2910	2.71	4.3	5.43	6.3	7.02	7.62	8.4	8.6	**9**
Root square	a=1.0001	1	2	3	4	5	6	7	8	**9**
Inverse linear*	-	1	1.13	1.29	1.5	1.8	2.25	3	4.5	**9**
Balanced*	-	1	1.22	1.5	1.86	2.33	3	4	5.67	**9**

*. Does not require this modification.

i.e. for geometric scales:

Table 8: Initial Information by Geometric Measurement Scale (Based on table 3)

Cri. / Alt.	*C1*	*C2*	*C3*	*C4*
A1	1	16	256	4
A2	64	4	1	256
A3	4	4	256	64
A4	16	64	4	16

Table 9: Modified Information (Based on table 7, for geometric measurement scale)

Cri. / Alt.	*C1*	*C2*	*C3*	*C4*
A1	1	16	256	4
A2	64	4	1	256
A3	4	4	256	64
A4	16	64	4	16

Table 10: AHP Result (For Information with Modified Geometric Measurement Scale)

Cri. / Alt.	C1 .176	C2 .187	C3 .333	C4 .304	Composite priorities
A1	.091	.257	.434	.091	.237
A2	.476	.148	.048	.475	.272
A3	.158	.148	.434	.275	.284
A4	.274	.446	.083	.158	.208

A comparison of the test results is given in table 11.

Table 11: AHP Results for Different Modified Measurement Scales (table 7)

Scale type	Priorities (rank and intensity)
linear	**A3 > A2 = A4 > A1** .275 .250 .250 .225
power	**A3 > A2 = A4 > A1** .275 .250 .250 .225
geometric	A3 > A2 > A1 > A4 .284 .272 .237 .208
logarithmic	A3 > A4 > A2 > A1 .267 .263 .241 .228
Root square	**A3 > A2 = A4 > A1** .275 .250 .250 .225
Inverse linear	A3 > A2 > A1 > A4 .292 .287 .263 .158
balanced	A3 > A2 > A1 > A4 .288 .279 .249 .183

Findings

1. As seen from the table, linear, power, and root square scale results (after modification) are the same, as compared with standard.
2. Another importance point to observe is that, in the power and root square modified scales, priorities (rank and intensity) are changed (from: A3, .289 > A2, .273 > A1, .227 > A4, .211; for power scale, and A4, .263 > A3, .262 > A2, .241 > A1, .233; for root square scale, to: A3, .275 > A2, .250 = A4, .250 > A1, .225), and similar to standards. Therefore, the power and root square scales to scale parameters (**a**) are highly sensitive and differ from other scales.
3. It is remarkable to observe that in this illustrative example, from all of the modified measurement scales, the A3 ranking is best.

Changing the Scale Parameters (a parameters)

In this section, the change of the scale parameters from a=2 (except for linear scale; from a=1), to a=3 and a=5 for all scales, is considered (table 12).

Table 12: The Change of Scale Parameter Results

Scale type	Scale parameters	values								
Linear	a=3	3	6	9	12	15	18	21	24	27
	a=5	5	10	15	20	25	30	35	40	45
Power	a=3	1	8	27	64	125	216	343	512	729
	a=5	1	32	243	1024	3125	7776	16808	32768	59049
Geometric	a=3	1	3	9	27	81	243	729	2187	6561
	a=5	1	5	25	125	625	3125	15625	78125	390625
Logarithmic	a=3	.631	1	1.262	1.465	1.631	1.771	1.893	2	2.096
	a=5	.431	.683	.861	1	1.113	1.209	1.292	1.365	1.431
Root square	a=3	1	1.260	1.442	1.587	1.710	1.817	1.913	2	2.080
	a=5	1	1.149	1.246	1.320	1.380	1.431	1.476	1.516	1.552
Inverse linear*	a=3	-	-	-	-	-	-	-	-	-
	a=5	-	-	-	-	-	-	-	-	-
Balanced*	a=3	-	-	-	-	-	-	-	-	-
	a=5	-	-	-	-	-	-	-	-	-

*. Will not be examined. Because, different parameters (from **a** parameters) were used.

A comparison of the test results is given in table 13.

Table 13: Changing the Scale Parameter Results

Scale type	*Scale parameters*	*Priorities (rank and intensity)*
Linear	a=3	**A3 > A2 = A4 > A1 .275 .250 .250 .225**
	a=5	**A3 > A2 = A4 > A1 .275 .250 .250 .225**
Power	a=3	A3 > A2 > A1 > A4 .302 .295 .237 .166
	a=5	A3 > A2 > A1 > A4 .321 .320 .262 .098
Geometric	a=3	A3 > A2 > A1 > A4 .330 .329 .300 .041
	a=5	A2 = A3 > A1 > A4 .333 .333 .321 .014
Logarithmic	a=3	A3 > A4 > A2 > A1 .267 .263 .241 .230
	a=5	A3 > A4 > A2 > A1 .267 .263 .241 .230
Root square	a=3	A3 > A4 > A2 > A1 .261 .257 .245 .237
	a=5	A3 > A4 > A2 > A1 .257 .256 .246 .241

Findings

1. The results indicate that, only the linear measurement scale priorities in both scale parameters (a=3, 5) are the same as compared with standard.
2. As seen from the table, the geometric measurement scale results are shown different priorities for a=3 and 5.
3. Another important point to be observed is that, the A3 is ranking best for all the scales.
4. As seen from the table, the logarithmic measurement scale's result are showing same priorities for a=3 and 5. However, do not exhibit same priorities with standard.

Changing the System Numbers

Here, the switch from: 1-3-5-7-9 to: 2-4-6-8-10 values, is considered. A comparison of the test result is given in table 14.

Table 14: Changing the System Numbers Results

Scale type	*Priorities (rank and intensity)*
linear	A3 > A2 = A4 > A1 .271 .250 .250 .229
power	A3 > A2 > A1 > A4 .285 .267 .227 .221
geometric	A3 > A2 > A1 > A4 .318 .316 .269 .097
logarithmic	A3 > A4 > A2 > A1 .262 .258 .244 .236
Root square	A3 > A4 > A2 > A1 .262 .255 .246 .236
Inverse linear*	-
Balanced*	-

*. Will not be examined. Because, different parameters (from **a** parameters) were used.

Notes: the standard priorities for new situation (system 2-4-6-8-10), calculated as: A3, .271 > A2, .250 = A4, .250 > A1, .229.

Finding

1. The standard and linear measurement scale priorities, exhibit the same ranking, with slightly different intensities from the previous system (1-3-5-7-9).
2. The geometric measurement scale; however, do not exhibit same priorities with standard. Nevertheless, are shown the same priorities as previous systems (table 6).

Conclusion

In this paper, we are focusing on the scale element alteration from the three aspects: 1. the limited scale upper bound (up to 9), 2. Changing the scale parameters (**a** parameters), and 3. Changing the system numbers (from 1, 3…9; to 2, 4…10) and their impacts on priorities in AHP. The major findings are as follow:

1. Test Criterion 1

Test criterion 1 is showing that, the linear, power, and root square modified scale results are the same, as compared

with the standard. Whereas, beforehand, only the linear measurement scale gave the correct answers.

2. Test Criterion 2

Test criterions 2 are showing that, only the linear measurement scale in both scale parameters (a=3, 5), are having the same results, as compared with standard.

3. Test Criterion 3

Test criterions 3 are showing that, with changing the system numbers, simultaneously the standard and linear measurement scale priorities changed. So that, the same ranking with different intensities was obtained.

Generally, the results have shown that, the linear measurement scales have the best performance (or stability) in compare to another scale.

References

Cox, M. A. A. (2009). Multidimensional scaling as an aid for the analytic network and analytic hierarchy process. *Journal of Data Science,* 7(2009), 381-396.

Forman, E. H., & Gass, S. I. (2001). The analytic hierarchy process-an exposition. *Operations Research,* 49, 469-486.

Ishizaka, A., & Labib, A. (2009). *Analytic hierarchy process and expert choice: Benefits and limitations", OR Insight*, 22(4), 201-220.

Ishizaka, A., Balkenborg, D., & Kaplan T. (2011). Influence of aggregation and measurement scale on ranking a compromise alternative in AHP. *The Journal of the Operational Research Society,* 62(4), 700-710.

Ji, P., & Jiang, R. (2003). Scale transitivity in the AHP. *The Journal of the Operational Research Society,* 54(8), 896-905.

Luce, R. D. (1997). Quantification and symmetry. *British Journal of Psychology*, 88, 395-398.

Monat, J. P. (2009). The benefits of global scaling in multi criteria decision analysis, *Judgment and Decision Making,* 4(6), 492-508.

Munshi J. (2014). Method for constructing likert scales, (April 2, 2014). Retrieved from http://ssrn.com/abstract=2419366 or http://dx.doi.org/10.2139/ssrn.2419366

Narens, L. (1981). On the scales of measurement. *Journal of Mathematical Psychology,* 24, 249-275.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin,* 99(2), 166-180.

Pöyhönen, M., Hämäläinen, R. P., & Salo, A. (1997). An experiment on the numerical modeling of verbal ratio statements, 6. *Journal of Multi Criteria Decision Analysis,* 6.

Ramanathan, R. (2001). A note on the use of the analytic hierarchy process for environmental impact assessment. *Journal of Environmental Management*, 63, 27-35.

Sato, Y. (2001). The impact on scaling on the pair wise comparison of the Analytic Hierarchy Process. ISAHP, 2001, Berns Switzerland, August 2-4, 421-430.

Saaty, T. L. (2000). *Fundamentals of decision making and priority theory*. RWS publication, 6.

Saaty, T. L. (2004). *Scales from measurement not measurement from scales*. MCDM 2004, Whistler, B.C., Canada, August 6-11, 2004.

Saaty T. L. (2005). The Analytic Hierarchy and Analytic Network Processes for the measurement of intangible criteria and for decision making", in Multi Criteria Decision Analysis: state of the art survey (Figueira et al., Eds), Kluwer academic publisher, 345-406.

Salo, A., & Hamalainen, R. (1997). On the measurement of preferences in the Analytic hierarchy Process. *Journal of Multi Criteria Decision Analysis,* 6, 309-319.

Shinohara et al., (2001). *Why not use the Entropy method for weight estimations*. proceedings-6th ISAHP 2001, Berns, Switzerland, August 2–4, 2001, 431-434.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science,* 103(2684).

Triantaphyllou et al., (1994). On the evaluation and application of different scales for quantifying pair wise comparisons in fuzzy sets. *Journal of Multi Criteria Decision Analysis,* 3, 133-155.

Triantaphyllou, E., Shu, B., Sanchez, S. N., & Ray, T. (1998). Multi criteria decision making: an operations research approach. *Encyclopedia of Electrical and Electronics Engineering,* John Wiley & Sons, Newyork, 15, 175-186.

Turskis, Z., Zavadskas, E. K., & Peldschus, F. (2009). Multi criteria optimization system for decision

making in construction design and management. *Engineering Economics,* 1(61), 1-17.

Vachajitpan, P. (2004). *Measurements scales and derivation of priorities in pair wise and group decision making*, MCDM 2004, Whistler, B.C., Canada, August 6–11, 2004, 1-6.

Wang, X. (2007). Study of ranking irregularities when evaluating alternatives by using some ELECTRE methods and a proposed new MCDM method based on regret and rejoicing", MSc. Thesis, Louisiana State University. USA.

Wedley W. C. (2001). *AHP answers to problems with known composite values,* ISAHP 2001, Berns Switzerland, August 2-4, 2001, pp. 551-560.

Wedley, W. C. (2007). AHP/ANP-where is natural zero? ISAHP 2007, VinaDel Mar, Chile, August 3-6, 2007, pp. 1-15.

Article can be accessed online at http://www.publishingindia.com

CDPSM: A New Optimized Progressive Big Data Analytics For Partial Cancer Data using Amazon EMR

Shyam Mohan J. S.*

Abstract

Identifying of symptoms and treating cancer requires a thorough investigation and research requiring analysis of multiple levels available (partial or full) cancer data. Cancer data is spread across multiple data sources and data warehouses which are decentralized and are in different locations. Therefore only half or partial data is available. Progressive analytics provide an efficient way for querying data from various data clusters where each cluster contains only a piece of the examined data. We propose an effective framework to perform analytics over the available cancer data say Cancer Data Progressive Sampling Model (CDPSM) built for partially available cancer data deployed on Amazon EMR. Through a large number of experiments, we reveal the advantages of the proposed model and give numerical results comparing them with a deterministic model. These results indicate that the proposed model can efficiently reduce the time for performing progressive data analytics over partial cancer data and maintaining the quality of the result at high levels.

Keyword: Big Data, Progressive Sampling

Introduction

The process of collecting, organizing and analyzing the data collected from various application domains like financial services, life sciences, mobile services, etc. is known as Big Data as most of the data is unstructured. The main aim of performing analytics is to discover patterns from hidden data sets and to provide meaningful information for effective decision making. Effective decision making will be successful by data driven by analytics-generated insights. Majority of the analytics are concerned with batch processing systems built on top of the Hadoop. Computing systems for big data generally fall into two major categories with regards to time constraint. They are:

1. Batch processing, in which large volumes of on-disk data with no time constraints (e.g., MapReduce and GraphLab) is analyzed.
2. In-memory streaming processing, where the data is analyzed in real-time or short period of time (e.g., Storm, SAMOA). Huang and Liu proposed that next-generation computing systems for big data analytics should be capable of providing good hardware and software to match between big data algorithms and the underlying computing and storage resources.

Challenges and Problem Statement

Currently there are many accessible data sources which provide information relating to any gene viz., mRNA or protein sequence. mRNA is estimated by the number of known sequences which is called Sequence Retrieval System (SRS). Majority of the medical data sources maintained by different organizations is updated frequently. For example, Nucleotide or protein sequences with the same emphasis are updated at different intervals with various benchmarks and standards and majority of the databases are outdated and contain irrelevant information. One of the challenges for cancer data is that the information stored is decentralized and is growing exponentially. For testing or for diagnosis of cancer data is a difficult task as the data is continuously updated viz., the cataloging and assessment behavior of dynamic biological regulation is incomplete. New categorical discoveries and their related information have to be constantly and progressively built onto any comprehensive content structures. In our work, we built

* Assistant Professor, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram, Tamil Nadu, India.
Email: jsshyammohan@kanchiuniv.ac.in

a tensor based framework over cancer data and perform progressive analytics from partial information obtained which is used for treating and diagnosis of cancer.

Background and Literature Survey

According to Human Genome Project estimates, the human genome DNA contains around 3.2 billion base of pairs distributed among twenty-three chromosomes translated to about a gigabyte of information. By adding gene data, X-ray and NMR spectroscopy data, the volume increases dramatically in gigabytes or petabytes.

Some of the data collected from various repository are shown in table 1 and also can be found in references.

Contribution

CDPSM- a New Progressive Analytics Model for partially available Cancer Data

A new progressive model for partially available cancer data is proposed called CDPSM (Cancer Data Progressive sampling model).Without loss of generality; we assume that the users can encode their own sampling data i.e., by dividing them into various tuples or clusters at various intervals. For managing cluster data, schedulers are responsible for executing queries where the data is split into number of pieces or clusters for query execution and thereby achieving data parallelism and effective query composition. By performing successive progressive analytics on samples, they get incrementally processed providing a significant performance benefit. Schedulers work on the statistical assumptions and hence don't require any user involvement. Introducing CDPSM into an existing relational engine is easy because majority of the data appears in text. Implementing it on unstructured data is a challenging task. The table below shows the input data (taken in numeric) with progressive intervals and we rely on partial data.

Table 2: Input Data with Progressive Intervals

Interval	*User*	*Ad*
(0,∞]	user 0	a0
[1,∞)	user 1	a1
[2,∞)	user 2	a2

Amazon EMR

For effective and quick processing of vast amounts of data, we use Amazon Elastic MapReduce (Amazon EMR) web service. Amazon EMR provides effective Hadoop framework for processing huge and vast amount of data and hence providing an easy, fast, and cost-effective method for dynamically scalable Amazon EC2 instances Amazon EMR effectively handles big data use cases, log analysis, etc.

Taking Amazon EMR to the Cloud with Hadoop MapR Distribution

The MapR Distribution for Hadoop makes it easy for provision and managing Hadoop in the AWS Cloud in Amazon Elastic MapReduce (Amazon EMR). MapR is used for real time handling of cancer data and when used across various health care organizations for performing suitable data analytics will lead to effective treatment and diagnosis of cancer data and thereby providing a best proven platform for Big Data platform. Fortune 100 and Web 2.0 companies have already started using MapR for their organizations.

Running Parallel Hadoop Jobs in CDPSM Amazon EMR Cluster Using AWS Data Pipeline

For every available partial input of cancer data, we set start and limit for the cluster data without loss of generality. The input is a multi-stage job generated by partially available cancer data or simply Hadoop jobs. Each job consists of partial input data files, a partitioning key (or mapper), and a progressive reducer. For progressive analytics, we take Stream Insight to process cluster data. The Hadoop jobs can be run in parallel in clusters using Amazon Web Service (AWS) pipeline in CDPSM Amazon Elastic MapReduce (Amazon EMR).Cluster utilization can be increased using EMR. A scheduler is run on the top of Amazon EMR clusters for monitoring Hadoop activities and is responsible for running Hadoop activities in the cluster by assigning them to specific queues. New data arriving during real time processing can be specified to core Amazon EMR nodes and are automatically assigned to EMR clusters.

Table 1: Data Collected from Various Repository

S. No.	Repository	Sequence or category of Data	Data (Apprx.)	Data Growth (Apprx.)
1.	Prism (Progressive sampling Model)	Encode progressive sampling strategy into the data by augmenting tuples with explicit progress intervals.	Suitable for progressive analytics on big data in the Cloud.	Uses pipelining techniques.
2.	Now	Uses MapReduce computation paradigm	Binary Large Object Data (Blob)	Batch Processing
3.	GenBank (As on December 2014)	Nucleic acid sequences	178 million	Doubling in size for every 15 months
4.	SWISS-PROT database	Protein sequences	18 million	Doubling in size for every 15 months
5.	InSiteOne	Offers data archiving, storage, and disaster-recovery solutions to the health-care industry in U.S.	4 billion medical images and 60 million clinical studies from 800 clinical sites	Increasing at an approximate rate of about 12% per year.
6.	ESG (Enterprise Storage Group)	Forecasting Medical image data	Grows at a rate of 35 percent per year	2.6 million terabytes(2014)

Mechanism of Optimizing Time for CDPSM

For every available partial input of cancer data, we set start and limit for the cluster data. The major question is when to start and when to stop the model. The scheduler is responsible for retrieving partial results from clusters. Initially, we assume that the progressive interval starts from 0. Optimal Stopping Theory (OST) is used to stop CDPSM and to find the optimal or best time results based on sequentially observed random variables where stopping time is defined as a random variable $T \in 0, 1,...,\infty$.

The partial data is considered as random independent variables. The problem mentioned for

the partial data can be found in which are considered as a finite or an infinite horizon. For finite horizon scenario, the scheduler has to respond for a specific time interval and for an infinite horizon, the scheduler receives partial data and takes final decision in optimal time.

CDPSM For Multi Stage Processing

Schedulers

Effective resource allocation and job prioritization for a Hadoop cluster in CDPSM is provided by scheduler. Schedulers are chosen based on the type of application. For our stated problem, we choose Capacity scheduler and default scheduler both used interchangeably. Capacity scheduler uses queues for Hadoop clusters. The Capacity Scheduler is designed to run the jobs in Hadoop environment in a multi-tenant cluster which allows maximizing the throughput and allows sharing a large cluster. It sets limit for ensuring initialized or pending applications from the users and ensures stability of the cluster. Each job in CDPSM Hadoop is converted into a set of map and reduces tasks. CDPSM Hadoop resources in cluster enable sharing depending on the computing needs.

The configuration of capacity scheduler is done by the following commands:

HADOOP_CONF_DIR/capacity-scheduler.xml

HADOOP_YARN_HOME/bin/yarn rmadmin -refreshQueues

Launching an CDPSM on Amazon EMR cluster

CDPSM can be launched on the top of Amazon EMR cluster with MapR version 4.0.2 from AWS Management Console. It supports many editions of the MapR like Community Edition (M3), Enterprise Database Edition (M7), etc. The algorithm below shows launching an Amazon cluster.

Algorithm: Launching data instances in EMR cluster

Input:# of Mappers, X={a,b,c,….},Y={1,2,3,…}

Output: instances of cancer data

begin

Id =Partial Cancer Data Cluster;

Type =Data Cluster;

Hadoop Version =0.20;

Keypair =key value;

For each master Instance Type = k1.xlarge do

If (core Instance Type==k1.small) then continue

Core Instance Count= 30;

If (task Instance Type==k1.small) then continue

task Instance Count=30

Do boot strap Action set from D: //elasticmapreduce/bootstrap-actions/configure-hadoop,arg1,arg2,arg3, to D3: //elasticmapreduce/bootstrap-actions/configure-hadoop/configure-other-stuff,arg1,arg2;

End.

Evaluation

We have a multi-stage job generated by cancer data which is a partial data. Each job consists of partial input data files, a partitioning key (or mapper), and a progressive reducer as stated in section 6. For progressive analytics, each job consists of a special reducer which uses Stream Insight to process cluster data. Amazon EMR supports deployment on a cluster of machines. Due to this interesting feature, we have considered Amazon EMR for performing progressive analytics for partially available cancer data.

Experimental Setup

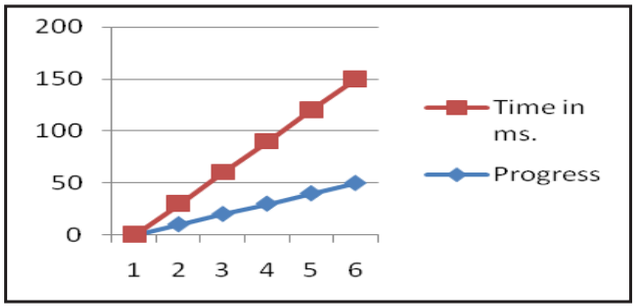
System Configuration

Machines in EMR are setup using Virtual Private Cloud (VPC) by changing the DNS Resolution and DNS hostname settings. Instances to communicate using EMR-managed security groups: The EC2 instance assigned is assumed to be default internal hostname. Input and output is stored in clusters. The cluster Id is known using VPC. Locally the system is of the configuration, 4GB RAM, and 1 TB of local storage, and 2Gbps allocated I/O bandwidth. We took 90 instances for our tests.

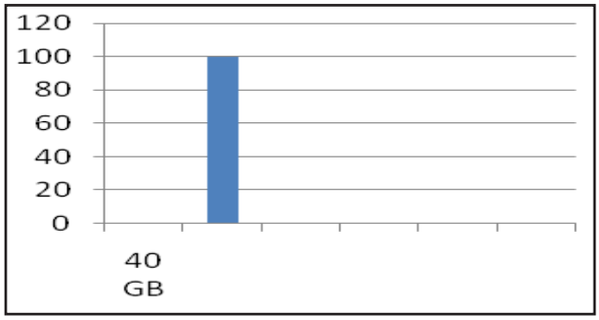
Datasets

We used the datasets available in for our evaluation based upon the aggregate amount of memory. We choose classification type as Brain cancer. Under this classification, we choose MicroRNA Data for Human Cancer data sets for performing analytics. Sample data sets are collected from over 299 patients that are available. We can even choose more samples based on the memory available. Factors in intergenic regions are also considered for data analysis. Input splits are created by shredding the data into various partitions with their corresponding Id. If the sample size is small we can rely on any algorithms like integer linear programming (ILP) algorithm. Stopping decisions are taken by Deterministic Stopping Model (DSM) which is considered for achieving the optimality. Missing or partial values of data can be found using parametric, non- parametric approach or Weibull distribution approach.

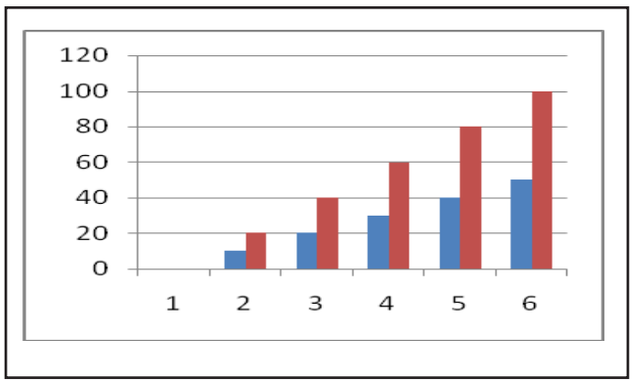
Figure a) shows the progressive computation for partially available cancer data (CDPSM) using Amazon EMR. Figure b) shows the performance of sample queries on Amazon EMR. Figure c) shows scalability of CDPSM for increasing data sets. Figure d) shows throughput of the machines.



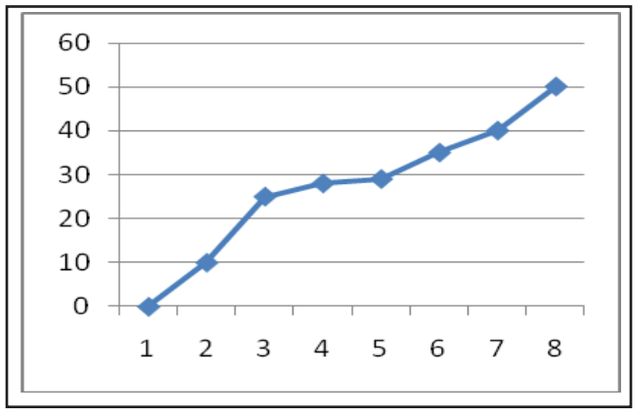
a) Progressive Computation-Time taken to process partial data in Amazon EMR.



b) Performance analysis of a sample query



c) Scalability with increase in data size



d) Throughput for number of machines

Fig. 1: Analysis of Various Data Sets on the Proposed Model

Conclusion

Progressive sampling on partial data is used to extract data for exploratory querying. Due to lack of proper tools for progressive analytics, obtaining results is a tedious task. We proposed a new progress model CDPSM deployed on Amazon EMR which allows efficient and deterministic query processing over partially collected sample data. For performing progressive sampling, we rely on Amazon EMR which provides a new framework for performing big data analytics over partially available cancer data where progress is achieved as a first-class citizen. Combination of all the above factors will lead to an effective and progressive big data analytics for partially available cancer data in optimized time.

References

Singh, S., & Singh, N. (2012). *Big data analytics,* In Proceedings of the International Conference on Communication, Information and Computing Technology.

Baldominos, A., Albacete, E., Saez, Y., & Isasi, P. (2014). *A scalable machine learning online service for big data real-time analysis.* In 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD), pp.1–8.

Huang, H., & Liu, H. (2014). *Big data machine learning and graph analytics: Current state and future challenges.* In 2014 IEEE International Conference on Big Data (Big Data), Oct. 2014, pp.16–17.

U.D. Energy, Insights learned from the human DNA sequence, what has been learned from analysis of the working draft sequence of the human genome? What is still unknown?, online, http://www.ornl.gov/hgmis, accessed on 2nd May 2011.

Hey, A. J., & Trefethen, A. E. (2003). The data deluge: An e-science perspective.

NCBI, Genebank statistics. Retrieved from http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html, accessed on 2nd Aug. 2014.

Uniprotkb/swiss-prot protein knowledgebase release 2011_04 statistics. Retrieved from http://expasy.org/sprot/relnotes/relstat.html, accessed on 10th April 2011.

Uniprotkb/trembl protein knowledgebase release 2011_04 statistics, online, http://www.ebi.ac.uk/uniprot/TrEMBLstats, accessed on 10th April 2011.

Baluja, T. Electronic patient records will soon end doctor's scrawl on paper, the globe and mail. Retrieved from http://www.theglobeandmail.com/news/national/toronto/electronic-patient-records-will-soon-end-doctors-scrawl-on-paper/article1982647

Insiteone official website. Retrieved from http://www.insiteone.com/, accessed on 10th April 2011.

Nuclear Cardiology Markets, TriMark Publications, LLC, 2007.

EMR, E. (n.d.). News, Dell launches new cloud-based services for hospitals and physician practices, online. Retrieved from http://www.emrandhipaa.com/news/2011/02/21/dell-launches-new-cloud-based-services-for-hospitals-and-physician-practice, accessed on 3rd April 2011.

Efficient Machine Learning for Big Data: A Review, Omar Y.Al-Jarrah, Paul D.Yoob, Sami Muhaidat, George K. Karagiannidis,KamalTahaa, http://dx.doi.org/10.1016/j.bdr.2015.04.001 Elsevier-2214-5796/2015.

Chandramouli, B., Goldstein, J., & Quamar, A. (2013). *Scalable Progressive Analytics on Big Data in the Cloud.* In the Proceedings of the VLDB Endowment, 6(14). Riva del Garda, Trento, Italy.

Chaudhuri, S., Das, G., & Srivastava, U. (2004). Effective use of block-level sampling in statistics estimation, in: SIGMOD.

https://aws.amazon.com/elasticmapreduce/

Peskir, G., & Shiryaev, A. (2006). *Optimal Stopping and Free Boundary Problems,* ETH Zuerich, Birkhäuser.

Kolomvatsos, K., Anagnostopoulos, C., & Hadjiefthymiades, S. (2015). An Efficient Time Optimized Scheme for Progressive Analytics in Big Data. Retrieved fromhttp://dx.doi.org/10.1016/j.bdr.2015.02.0012214-5796, Big Data Research, February-2015, Elsevier.

https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/CapacityScheduler.html.

http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi

Guidelines for Authors

International Journal of Business Analytics and Intelligence welcomes original manuscripts from academic researchers and business practitioners on the topics related to descriptive, predictive and prescriptive analytics in business. The authors are also encouraged to submit perspectives and commentaries on business analytics, cases on managerial applications of analytics, book reviews, published-research paper reviews and analytics software reviews based on below mentioned guidelines:

Journal follows online submission for peer review process. Authors are required to submit manuscript online at http://manuscript.publishingindia.com

Title: Title should not exceed more than 12 Words

Abstract: The abstract should be limited to 150 to 250 words. It should state research objective(s), research methods used, findings, managerial implications and original contribution to the existing body of knowledge

Keywords: Includes 4–8 primary keywords which represent the topic of the manuscript

Main Text: Text should be within 4000-7000 words Authors' identifying information should not appear anywhere within the main document file. Please do not add any headers/footers on each page except page number. Headings should be text only (not numbered).

Primary Heading: Centered, capitalized, and italicized.

Seconday Heading: Left justified with title-style capitalization (first letter of each word) and italicized.

Tertiary Heading: Left justified and indented with sentence-style capitalization (first word only) in italics.

Equations: Equations should be centered on the page. If equations are numbered, type the number in parentheses flush with the left margin. Please avoid using Equation Editor for simple in-line mathematical copy, symbols, and equations. Type these in Word instead, using the "Symbol" function when necessary.

References: References begin on a separate page at the end of paper and arranged alphabetically by the first author's last name. Only references cited within the text are included. The list should include only work the author/s has cited. The authors should strictly follow APA style developed by American Psychological Association available at American Psychological Association. (2009). Publication manual of the American Psychological Association (6th Ed.). Washington, DC.

Style Check

To make the copyediting process more efficient, we ask that you please make sure your manuscript conforms to the following style points:

Make sure the text throughout the paper is 12-point font, double-spaced. This also applies to references.

Do not italicize equations, Greek characters, R-square, and so forth. Italics are only used on p-values.

Do not use Equation Editor for simple math functions, Greek characters, etc. Instead, use the Symbol font for special characters.

Place tables and figures within the text with titles above the tables and figures. Do not place them sequentially at the end of the text. Tables and figures must also be provided in their original format.

Use of footnotes is not allowed; please include all information in the body of the text.

Permissions

Prior to article submission, authors should obtain all permissions to use any content if it is not originally created by them.

When reproducing tables, figures or excerpts from another source, it is expected to obtain the necessary written permission in advance from any third party owners of copyright for the use in print and electronic formats. Authors should not assume that any content which is freely available on the web is free to use. Website should be checked for details of copyright holder(s) to seek permission for resuing the web content

Review Process

Each submitted manuscript is reviewed first by the chief editor and, if it is found relevant to the scope of the journal, editor sends it two independent referees for double blind peer review process. After review, the manuscript will be sent back to authors for minor or major revisions. The final decision about publication of manuscript will be a collective decision based on the recommendations of reviewers and editorial board members

Online Submission Process

Journal follows online submission for peer review process. Authors are required to register themselves at http:// manuscript. publishingindia.com prior to submitting the manuscript. This will help authors in keeping track of their submitted research work. Steps for submission is as follows:

1. Log-on to above mentioned URL and register yourself with "International Journal of Business Analytics & Information"
2. Do remember to select yourself as "Author" at the bottom of registration page before submitting.
3. Once registered, log on with your selected Username and Password.
4. Click "New submission" from your account and follow the 5 step submission process.
5. Main document will be uploaded at step 2. Author and Co-author(s) names and affiliation can be mentioned at step 3. Any other file can be uploaded at step 4 of submission process.

Editorial Contact
Dr. Tuhin Chattopadhyay
Email: dr.tuhin.chattopadhyay@gmail.com
Ring: 91-9250674214

Online Manuscript Submission Contact
Puneet Rawal
Email: puneet@publishingindia.com
Ring: 91-9899775880

International Journal of Business Analytics and Intelligence

SUBSCRIPTION DETAILS

Dispatch Address:-
The Manager,
International Journal of Business Analytics and Intelligence
Plot No-56, 1st Floor
Deepali Enclave, Pitampura
New Delhi -110034
Ph - 9899775880

Subscription Amount for Year 2017

	Print	Print + Online
Indian Region	Rs 2700	Rs 3400
International	USD 150	USD 180

Price mentioned is for Academic Institutions & Individual. Pricing for Corporate available on request. Price is Subject to change without prior notice.

Payment can be made through D.D./at par cheque in favour of "Publishing India Group" payable at New Delhi and send to above mentioned address.

Disclaimer

The views expressed in the Journal are of authors. Publisher, Editor or Editorial Team cannot be held responsible for errors or any consequences arising from the use of Information contained herein. While care has been taken to ensure the authenticity of the published material, still publisher accepts no responsibility for their inaccuracy.

Journal Printed at Anvi Composers, Paschim Vihar.

Copyright

www.manuscript.publishingindia.com

PublishingIndia.com

Publishing India Group

Plot No. 56, 1st Floor, Deepali Enclave
Pitampura, New Delhi-110034, India
Tel.: 011-47044510, 011-28082485
Email: info@publishingindia.com
Website: www.publishingindia.com

ISSN 2321-1857
01
9 772321 185001