Cairo University
Faculty of Graduate Studies for Statistical Research

The 54th Annual Conference on Statistics, Computer Sciences and Operations Research

Applied Statistics

9-11 Dec. 2019

Index
APPLIED STATISTICS

1	**Marketing Efficiency of Tomato in Sahl El Tina Area** Soha M.Eldeeb and Dalia E.Abozied	**1-9**
2	**Optimal Using of Irrigation Water Resources in Desert Areas Case Study for Some Medicinal and Aromatic Plants Production** Dalia E.Abozied	**10-18**
3	**Bayesian ldentification Methods for Autoregressive Models: A Comparison Study** Ayman A.Amin	**19-28**
4	**Robust Generalized Minimum Variance Control (GMVC): An Application on Economic Time Series** Zidan,A.I.	**29-44**
5	**Chaotic Analysis of Egyptian Stock Market with the Application to EGX 30 Price Index** Eman Talaat Morad , El-Houssainy A. Rady and Ayman A. Amin	**45-62**
6	**New Shrinkage Parameters for Liu-type Zero Inflated Negative Binomial Estimator** El-Houssainy A. Rady, Mohamed R. Abonazel and Ibrahim M.Taha	**63-84**
7	**Intra Plot Variation and Multivariate Analysis in Maize Fertilization Experiments** Abdalla, A.F.M, Okaz, A.M.A,Hager, M.A and Hasham, M.M.	**85-98**
8	**An Overview of Composite Indices: A Dimension Reduction Perspective** Asmaa Mohamed Sayed Abdelatif, and Yasmin Mohamed Ibrahim	**99-113**
9	**The Change in the Patterns of the Egyptian Family Expenditure in Health (2005-2013)** Amani Musa Mohammed and Shimaa Farouk Hassan	**114-130**
10	**An Econometric Analysis of Modernized Agriculture in Reducing Multidimensional Poverty in Uganda** Emong Herbert Robert	**131-148**

Marketing Efficiency of Tomato in Sahl El Tina Area

Soha M. Eldeeb and Dalia E. Abozied[1]

Abstract

The paper aimed to estimate the marketing margin, marketing efficiency, producer share from the consumer prices in L.E, also concentrate on factors affect market efficiency, identify constraints that effect marketing of the tomato**.** The paper was conducted in Sahl El Tina area, 123 farmers were selected randomly present about 25% from all tomato farmers in the area. Data primary survey was conducted to elicit the required information of marketing channel. Results indicated that, the cultivated area was decreased by 13.1 feddan annually by decrease rate reached 3% annually and the production was decreased by 60.03 ton annually by decrease rate reached 3% annually , an estimate was made to find out the producer's share from the consumer price in L.E. reached about 43.7% this percentage was low compared to the intermediary share which reached 56.3% , while market efficiency reached 26.2% which mean decreasing in the market efficiency of tomato products in the studied area, this can be due to high marketing costs and the perishability of tomato, The marketing efficiency of Tomato was found influenced by several significant parameters like marketing cost(these factors negatively influenced the market efficiency) while volume of the produce handled (positively affected the market efficiency).

Key words: Marketing margin, Price spread, Wholesaler, Retailer, Marketing channel

Introduction

Marketing plays a critical role in meeting the overall goals of food security, poverty alleviation and sustainable agriculture, particularly among smallholder farmers in developing countries (Altshul, 1998; Lyster, 1990). Marketing by smallholder farmers was constrained by poor infrastructure, distance from the market, lack of assets (for example lack of own vehicles) and inadequate market information. Lack of bargaining power along with various credit bound relationships with the buyers has led to farmers being exploited during the transaction where most of the farmers become price takers. The majority of the farmers are smallholders and hence unable to obtain a fair price for their produce. This results to farmers not being able to sustain their livelihood. The structure of the traditional vegetable supply chains is such that there are a large number of intermediaries (e.g. vegetable collectors, transporting agents, commission agents etc.) between the producer and the consumer. Addition of the marketing margins of all these intermediaries coupled with almost 30 to 40 percent of the vegetables being wasted as post harvest losses have eventually resulted in producers receiving a very low price for their produce while at the other end the consumers are compelled to pay a highly inflated price for their purchases

[1]Assistant professor, Agricultural Economic Department, Socio-Economic Division, Desert Research Center, Cairo. Egypt

(Hettige & Senanayake, 1992; Kodithuwakku, 2000). Markets tend to be disorganized when the farmers and traders who do not fully rely on their vegetable business for a steady income often sell their produce at almost any price offered. Retail agents often encourage this since it provides an opportunity for them to make more profits. In the long run, this is not good for the industry since it promotes an erratic supply and unrealistic pricing structure. Marketing information is important in assisting growers at crop planning stage before planting and to sell Surplus produce. In the absence of such marketing information the retail end of the industry does not respond to supply and demand and pricing is set artificially, and it remains static. The tomato crop plays a major nutritional and economic role in Egypt's vegetable crop production, and in spite of all this, it is affected by the economic and political changes as well as natural and climatic conditions causing productivity and price fluctuations as well as a high percentage of marketing losses, which make it difficult to organize the supply and achieve production and marketing efficiency, which targeted for increased domestic and external demand. So the research aimed to attain the proposals that will benefit decision makers and to find out proper production and marketing policy, and therefore, benefit the farmers, consumers and the state. (M. Bediwy, 2016). Tomato is a perishable product that needs more efficient production and marketing activities (Michele, 1996).

Study Area

Sahl El Tina was selected to represent marginal ecosystem. The irrigation water was obtained from mixed water (Nile water + drainage waters) of El Salam Canal. The soils are characterized by severe salt affected, differ in depth and stratified profile layers. Total cultivated area reach about 35.1 thousand feddan (about 70% of the total area in the region) while the infrastructural occupied about 14.9 thousand feddan (about 30% of the total area in the region).Sahl El Tina composed from seven villages. The area cultivate mainly traditional crops like (wheat, sugar beet, alfalfa, barley, vegetables) this traditional cropping pattern can be related to the characteristic of the beneficiaries of those villages, since most of the beneficiaries are from old valley and delta governorates which are famous by cultivating this type of cropping pattern

Study problem

Indeed tomato get a high economic and nutrition importance. While the economic and climate changes conditions and high rates of losses during marketing can cause prices fluctuations thus the supply cannot meet the local or export demand. This can affect market efficiency in a negative way.

Objectives

The objectives are:

- Estimation of the marketing margin, price spread, marketing efficiency and producer share from the consumer prices in L.E in the marketing channel.

- Study of factors influencing the marketing efficiency.
- Identify the constraints faced Farmers in marketing of tomato

Data and Methods

The research was conducted in Sahl El Tina area, research depended on two sources of data secondary and primary data, 123 farmers were selected randomly present about 25% from all tomato farmers in the area which reached 482 farmers. Data primary survey was conducted to elicit the required information of marketing channel, market margin, price spread, producer share in consumer pound and constraints, Shepherd formula, Acharya modified marketing Efficiency formula were used for estimating marketing efficiency.

Marketing Margin analysis

Marketing margin analysis was used to estimate the margin in terms of revenue and profit that accrue to the tomato marketers.

Marketing margin can be estimated as:

- Marketing Margin = Selling Price – Cost Price
- Net marketing margin = Marketing margin–marketing cost.
- Net Marketing Margin for Wholesaler = Wholesale marketing margin – Wholesale marketing cost.
- Net Marketing Margin for Retailer = Retailer marketing margin – Retailer marketing cost.

Marketing Efficiency

The Marketing efficiency is measured as:

- Marketing efficiency = (Gross marketing margin/ marketing cost) × 100

Also, Sherpherd formula technique was used as follows:

- Marketing efficiency = (Consumer price/Total marketing cost) – 1

Price Spread

It is the difference between the two prices, i.e., the price paid by the consumer and the price received by the producer.

For e.g. P1-P2, Where, P1 is price at one level or stage in the market, P2 is price at another level

Factors Affecting Marketing Efficiency

Multiple linear regression analysis with following variables was done to know the effect of these variables on marketing efficiency.

$Y=f(X_1, \ldots\ldots\ldots, Xn)$ Where, Y=Marketing efficiency (%)، X1 = Marketing cost (LE.) ، X2 = Marketing margin (LE) ، X3 = Transport cost (LE) ، X4 = Labor wages (LE.) ،X5 = Volume of produce handled(kg) ، X6 = Length of the market channel (No. of market intermediaries)

Results and Discussion

Table (1) and Figure (1) show that, developing the cultiveted area and production of tomato in Sahl El Tina area ,the cultiveted area and production of tomato was changed from year to another during (2009 -2018). Average cultiveted area of tomato reached about 401 feddan, the cultivated area was decreased by 13.1 feddan annually and was decreased with a decreasing rate reached 3% annually. For Tomato production the average production was about 2207 ton , the production was decreased by 60.03 ton annually and was decreased with a decreasing rate reached 3% annually. Decreasing tomato cultivated area might be due to the security status in the area which causes transportation difficulties to the markets which reflect a high damage rate because of tomato nature and it causes big loses for tomato farmers, for this farmers may be avoiding tomato Cultivation.

Table1. Developing the Cultivated Area and production of Tomato in Sahl El Tina Area during (2009 -2018)

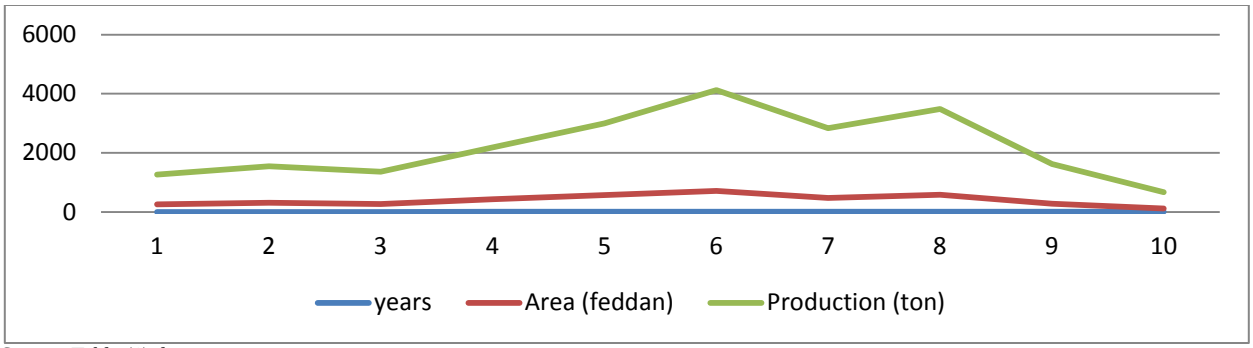
years	Area (feddan)	Production (ton)
2009	252.2	1268.5
2010	308.8	1544
2011	273.3	1366.5
2012	436.1	2180.5
2013	570.1	2993
2014	716.1	4117.5
2015	472.5	2835
2016	580.5	3483
2017	281.5	1618.6
2018	121.5	668.2
Average	401.3	2207.5
Change amount	- 13.1.	- 60.03.
Change rate	- 0.03	- 0..03

Source: Agriculture department ,Sahl El Tina ,different issues

Change amount = (last year-first year)/number of years

Change rate = change amount / average

Figure1. Developing the Cultivated Area and ~~pr~~oduction of Tomato in Sahl El Tina Area during (2009 -2018)



Source: Table (1) data

Market Channel

Tomato marketing channel follows in the study area is shown in (Figure 2). The principal agents in the tomato marketing channel were; producer, commission agent, wholesaler and retailer. The results of the study are discussed below:

1. **Producer**: Tomato cultivation is carried out by a large number of farmers who are geographically dispersed in various locations. It was found that the tomato farmers sold tomato in the local markets and to wholesale market (Obour and Ismailia markets) (Figure 2).

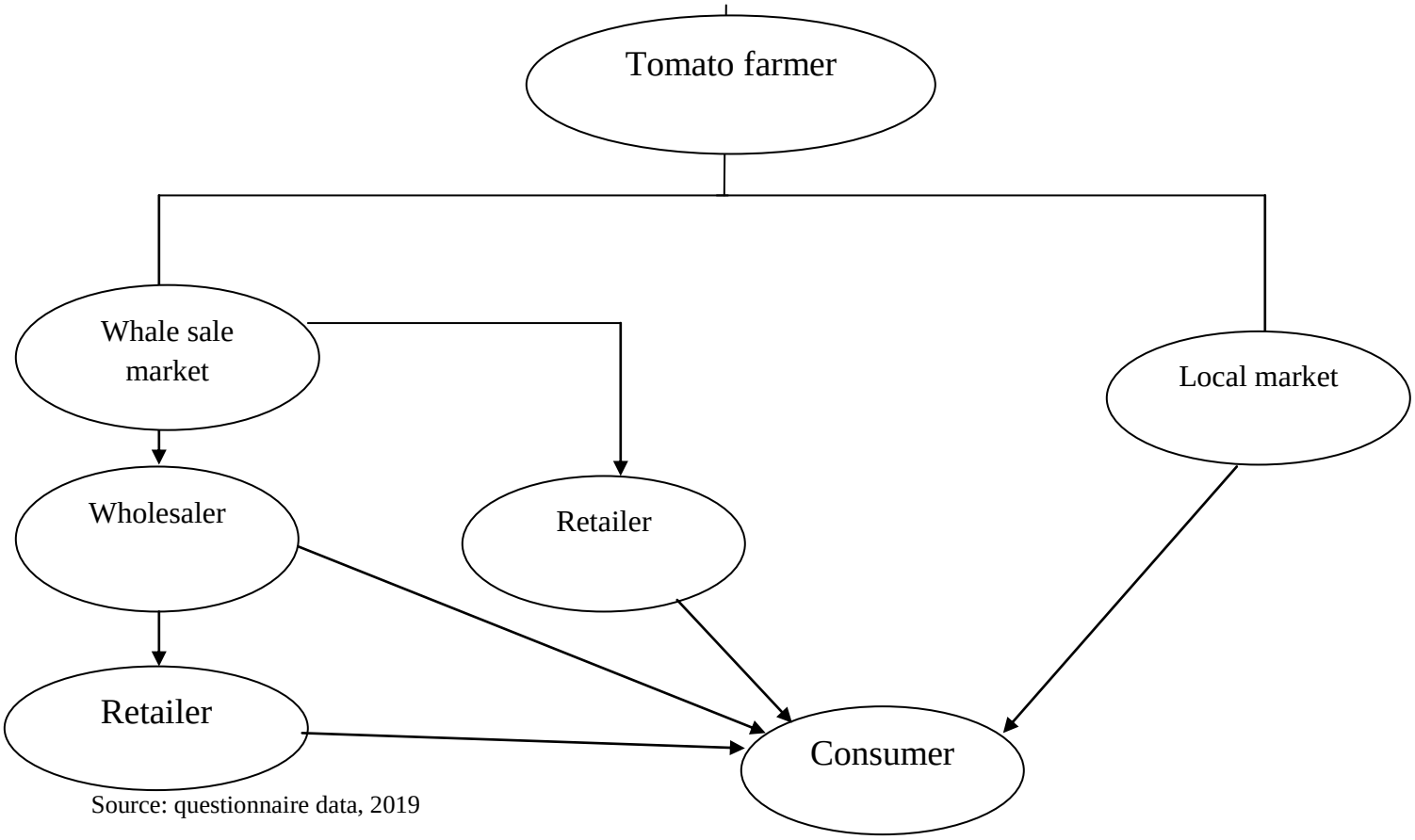
2. **Commission Agent:** The main function of commission agents is to bring buyers and sellers together. Commission agents maintain contacts with inter-regional wholesale markets and possess comprehensive and accurate information. The commission agent is the principal intermediary around which all marketing activities rotate. Commission agents perform activities on commission reached 10%* from the farmer and do not accept any title of goods while simply selling the produce brought by producers or contractors, commission agents have establishments in the markets equipped with telephone and other communication facilities.

3. **Wholesalers:** wholesalers buy and sell large quantities of farm products. Usually wholesalers perform business in wholesale markets. Wholesalers deal in several commodities such as fruits; vegetables and other agricultural produce within inter-regional markets and also supply produce to processing industries, exporters and retailers according to their demand. Wholesalers maintain contacts with commission agents in wholesale markets and retailers in the local area. A wholesaler usually purchases tomato from the commission agent at open auction and sells smaller quantities to the retailers and consumers.

4**. Retailers**: Market activities come to an end with the retailer before the product reaches in the hands of final consumer, retailers buy and sell small quantities according to the demand of consumers in the area and maintain direct contact with consumers and make transactions according to the qualitative and quantitative aspects of the products. A small number of tomato retailers occupy small shops in the main vegetable markets or in the town. The majority of tomato retailers are hawkers, selling from barrows consisting of a wooden platform mounted on four wheels, retailers move from street to street to offer tomato for sale. Among tomato retailers there is a high degree of competition. Retailers buy tomato from the wholesaler and sold to consumers. Consumers received rest directly from

wholesaler. Similarly, retailer supplied another of market tomato to consumers that he bought directly in the wholesale market (Figure 2).

Figure 2. Market Channel for Tomato in Sahl El Tina area



Source: questionnaire data, 2019

Farm, wholesale, retail price and consumer pound distribution

Table (2) shows that , farm price reached about 3500 LE/ton ,wholesale price reached about 6000 LE/ton and the retail price reached about 8000 LE/ton, producer share reach about 43.7% of the consumer price what is lower than the intermediary share which reach about 56.3% in the area as shown in table (2)

Table2. Farm, Wholesale, Retail price and Consumer Pound Distribution

Prices			Consumer pound distribution			Intermediary share (7)
Farm price(1)	Wholesale price(2)	Retail price(3)	Farmer share (4)	Wholesaler share (5)	Retail share (6)	
3500	6000	8000	43.7	31.3	25	56.3

(4)= (1/3)*100 (5) = (2-1/3)*100 (6) = (3-2/3)*100 (7) = (5+6)

Source: Collected and calculated from the questionnaire, (April) 2019

Marketing Margin of Tomato Traders.

Table (3) summarizes and presents the marketing margin for the wholesalers and retailers. This table indicates that the marketing margin between the wholesaler and the farmer reached about 2500 LE /ton, the marketing margin between the retailer and the wholesaler reached about 2000 LE /ton , the marketing margin between the farmer and the retailer reached about 4500 LE /ton also table (3) shows that, market efficiency reached 26.2% which mean decreasing the market efficiency of tomato because high marketing cost of tomato also decreasing in the marketing efficiency of tomato might be due to the perishability of tomato, increasing in tomato price fluctuations and the monopoly of major traders in tomato markets, especially in wholesale markets during shortage times of production, which lead to increase the profits of intermediaries in the marketing process and decrease the price of the product which decrease marketing efficiency

Table 3. Marketing Margins and Coefficient of Market Efficiency

Marketing Margins			Total marketing margins	*Coefficient of market efficiency
Farm - Wholesale	Retail - Wholesale	Retail - Farm	4500	
2500	2000	4500		26.2

*Coefficient of market efficiency = {100-(1600/1600+4500)}*100

Source: Collected and calculated from the questionnaire, (April) 2019

Factors Affecting Marketing Efficiency

The marketing efficiency of Tomato was found influenced by several significant parameters as shown in Table (4). Marketing cost, marketing margin, transport cost negatively influenced and volume of the produce handled positively affected.

Table4. Factors Affecting Significantly on Market Efficiency of Tomato

Factors	Coefficient value
Marketing cost	-0.38*
Marketing margin	-0.10*
Transport cost	-0.06*
Volume handled	0.93**

*Significant at 5% level of significance, **Significant at 1% level of significance

Source: Collected and calculated from the questionnaire, (April) 2019

Constraints Faced Farmers in Marketing of Tomato

Table (5) presents that, the constrains faced by farmer in marketing tomato were high marketing costs by 99%, high cost and difficulties of transportation by 98%, low producer price by 97%, Labor shortage by 95%, long distance to the markets by 94%, Facing difficulties in selling product by 92%, limited access to other markets by 91%, damage

during transportation by 90%, Limited market information by 89%, price fluctuations by 88% and unsuitable market infrastructure by 87% .

Table5: Constraints Faced Farmers in Marketing of Tomato

No	Constraints	%	rank
1	High marketing costs	99	1
2	low producer price	97	3
3	Long distance to the markets	94	5
4	Limited market information	89	9
5	Facing difficulties in selling product	92	6
6	Limited access to other markets	91	7
7	Unsuitable market infrastructure	87	11
8	Labor shortage	95	4
9	Price fluctuations	88	10
10	High cost of transportation	98	2
11	Damage during transportation	90	8

Source: Collected and calculated from the questionnaire, (April) 2019

References

Altshul, H. (1998). Output to purpose review of DFID's Crop Post-harvest Programme, Value Addition to Agricultural Products. Natural Resources International Symposium, 53-61p

Baki-Abdul, A. Aref , John Teasdale. (1997) Sustainable Production of Fresh Market Tomatoes andOther Summer Vegetables with Organic Mulches, Farmers' Bulletin No 2279. USDA-Agriculture Research Service,Washington, D.C. 23 p.

Hettige, S. T. Senanayake, S. M. P. (1992). Highland Vegetable Production and Marketing Systems. A report prepared for Agriculture Cooperative Development International/USAID

Kohls, R, Uhl, J. (1985) Marketing of Agricultural Products. New York: Mac Milan. Sixth Edition.

Kodithuwakku, K. A. S. S. (2000). Analysis of Tomato Supply Chains in the Kandy istrict. An unpublished report prepared for University of Peradeniya

Lyster, D. M. (1990). Agricultural marketing in KwaZulu: a farm-household perspective. Unpublished MSc Agric thesis, University of Natal, Pietermaritzburg.

Tahir Zahoor Chohan , Sarfraz Ahmad.(2008). Post-Harvest Technologies and Marketing Channel in Tomato Production in Danna Katchely, Azad Jammu Kashmir. Pakistan Journal of life social science, 6(2), 80-85p.

الكفاءة التسويقية للطماطم فى منطقة سهل الطينة

سهى مصطفى الديب داليا السيد ابو زيد

تعد الطماطم واحدة من أكثر الخضروات انتشارًا على نطاق واسع في العالم ، كما أنها تعتبر من محاصيل الخضر التى لها دورًا غذائيًا واقتصاديًا كبيرًا,الا أنها تتاثر بالتغيرات الاقتصادية والسياسية كما تتأثر بالظروف الطبيعية والمناخية والتي تتسبب في التقلبات الإنتاجية والسعرية بالإضافة إلى ارتفاع نسبة الفاقد التسويقى منها مما يجعل من الصعب تنظيم المعروض وتحقيق الكفاءة الانتاجية والتسويقية ،ايضا الطماطم منتج سريع العطب يحتاج إلى أنشطة إنتاجية وتسويقية أكثر كفاءة لذلك يهدف البحث إلى تقدير الهوامش التسويقية والانتشار السعرى والكفاءة التسويقية ونصيب المزارع في جنية المستهلك فى المسلك التسويقى للطماطم ، وايضا دراسة العوامل التي تؤثر على الكفاءة التسويقية وتحديد المشكلات التي تواجه المزارعين في تسويق الطماطم ، أجري البحث في منطقة سهل الطينة وقد اعتمد البحث على البيانات الثانوية والبيانات الاولية والتى تم تجميعها من 123 مزارع تم اختيارهم عشوائيا بنسبة تبلغ نحو 25% من اجمالى عدد مزارعى الطماطم والذى يبلغ عددهم نحو 482 مزارع وذلك باستخدام استمارة استبيان أعدت خصيصا لهذا الغرض، أشارت النتائج إلى أن متوسط المساحة المنزرعة من الطماطم خلال الفترة (2009–2018) بلغ نحو 401 فدان ، وانخفضت المساحة المزروعة بنسبة 13.1 فدان سنويًا بمعدل بلغ 3٪ سنويًا كما بلغ متوسط انتاج الطماطم خلال نفس الفترة حوالي 2207 طنً ، انخفض هذا الانتاج بمقدار 60.03 طن سنويًا بمعدل بلغ 3٪ سنويًا ، بلغ سعر المزرعة حوالي 3500 جنيه للطن ، وبلغ سعر الجملة حوالي 6000 جنيه للطن كما بلغ سعر التجزئة حوالي 8000 جنيه للطن ، تم تقدير نصيب المنتج في جنية المستهلك بنسبة 43.7٪ وهي نسبة منخفضة مقارنة بنصيب الوسطاء الذى بلغ 56.3٪ ، وبلغ الهامش التسويقى بين تاجر التجزئة وتاجر الجملة حوالي 2000 جنيه للطن كما بلغ الهامش التسويقى بين المزارع وتاجر التجزئة حوالي 4500 جنيه للطن ، اوضحت النتائج ايضا أن الكفاءة التسويقية بلغت 26.2 ٪ مما يعني انخفاض الكفاءة التسويقة للطماطم وذلك راجع الى ارتفاع التكاليف التسويقية ،ايضا اتضح من النتائج أن الكفاءة التسويقية للطماطم تتأثر بعدة عوامل منها مايؤثر سلبيا مثل التكاليف التسويقية ، الهوامش التسويقية ، تكلفة النقل ومنها مايؤثر ايجابيا مثل حجم المنتج ، وأخيرا وجد ان أهم المشاكل التي يواجهها المزارع في تسويق الطماطم هي ارتفاع التكاليف التسويقية ، ارتفاع تكلفة النقل ، انخفاض سعر الإنتاج ، نقص العمالة ، المسافة الطويلة إلى الأسواق ، مواجهة صعوبات في بيع المنتج ، محدودية النفاذ إلى الأسواق الأخرى ، الضرر أثناء النقل ، المعلومات التسويقية المحدودة ، تقلبات الأسعار والبنية التحتية الغير مناسبة للسوق على التوالي

الكلمات المفتاحية :الهوامش التسويقية ،الانتشار السعرى ، تاجر الجملة ،تاجر التجزئة ، المسلك التسويقى

Optimal Using of Irrigation Water Resources in Desert Areas Case Study for some Medicinal and Aromatic Plants Production

Dr. Dalia Abozaid*

Abstract

The agricultural sector is one of the most important consuming sectors which water consumes about 87% of the total actual in 2017[1]. irrigation water is considered the strategic element in agriculture, and base of agricultural expansion, so the problem of research was the limited water resources and the low level of efficiency used for irrigation, so research aim to estimate the economic efficiency and productivity for three types of irrigation water for medicinal and aromatic Plants crops in the Egyptian Agriculture.

The research was adopted to achieve its objectives on the use of design experiments analysis and used some economic efficiency indicators of irrigation water, through the assistance of the availability of published and unpublished data and issued by the scientific community. Search results in terms of indicators of economic efficiency of use of three irrigation water types have shown that, the low efficiency of the mint and basil plants in using treated agriculture drainage water for irrigation due to the decrease in the net yield of the water unit, which amounts to about 4789 and 5220.5 pounds, respectively. the average production of the green basil sample irrigated with blended irrigation water was 97.2% compared to the average production of the green basil sample irrigated with Nile water. The average of the green mint sample irrigated with mixed irrigation water was 96.4% compared to the average weight of the green mint sample irrigated with Nile water.

Introduction:

Medicinal and aromatic plants are unconventional crops, used by humans throughout the ages for various purposes. Used as a spice when cooking foods, other as medicine, and recently demonstrated the importance of medicinal and aromatic plants in treatment of many diseases that affect mankind, as they enter in many food industries as preservatives, taste and appetizer, in addition to the new ones are consumed in the form of stimulant or palliative drinks (Mamdouh Fathi,2003).

All definitions of medicinal and aromatic plants have agreed on the following: (Yasser Hamdy,2010) First: A medicinal plant contains one or more of its various organs containing one or more chemicals with low or high concentration, which has a physiological ability to treat a particular disease or at least reduce symptoms of this disease, either in its pure form after being extracted from the plant material or if it is used and is still in the first place in the form of fresh grass, dried or partially extracted. Second: An aromatic plant is defined as a plant that contains in one or more of its plant members or its modifications the volatile essential oils, whether in their own free form or in another form that transforms or hydrolyses into volatile essential oils of acceptable odor. Hence, the importance of these plants comes from an economic point of view as they have several uses that increase their economic value.

* Assistant professor, Agricultural Economic Department, Socio-Economic Division, Desert Research Center, Cairo. Egypt

Due to the decreasing availability of areas for cultivation of these plants, especially in the traditional cultivation areas in the old valley and the Delta, due to the unequal competition between medicinal and aromatic plants and strategic crops such as wheat, despite the increasing demand of them internally and externally.

So, moving out from the narrow valley to other areas more extensive for cultivation of these plants like the reclaimed lands in the desert areas or in the desert back yard of some old valley governorates. The advantages of these plants, in the low cost of production operations, ease of post-harvest transactions, relatively to long-term profitability compared to vegetables and fruit crops, as well as ease of marketing locally and globally .While productivity reached about 27 thousand tons, vary between fennel, anise, cumin, mint, marjoram, pond, lemon grass and thyme [4]. About 90-95% of the production of these plants annually exported. Revenue of these crops reached about $ 6.136 million during the 2017. Egypt's exports of medicinal and aromatic plants rank 11th on the list of exporting countries, occupying about 2.23% of the global market, Egypt's exports of these plants succeeded entering nearly 40 international markets [5]. France is the leading country in the import markets followed by United States of America, most important plants required for export, basil, mint, fennel, caraway, coriander and chamomile.

The continuous increase in the earth's population requires increasing quantities of water for agricultural needs. The progressive requirement for more water to irrigate crops for food when water resources are limited has lead to reuse and recycling of the available water in agriculture [6]; (Ragab R, 1997). In many regions of the world, agriculture drainage water is already used successfully for irrigation even when the water is saline [Grattan SR,et al 1994]. Irrigation with saline water has become necessary not only in parts of the world with limited supplies of good quality water but also in areas affected by shallow ground water.

Basil and mint are among the most exported plants of European countries, occupying an advanced position in the ranks of plants necessary for the domestic and international market .One of the most important characteristics of these plants, its resistance to salinity [Oudhia, P, 2003]. On this basis, the effect of three types of water on the production of these plants were studied: Nile water, mixed water (Nile water, 1: 1 treated agriculture drainage water with salinity < 3000 ppm), agriculture drainage water. Therefore, to assess the economic impact of these species on the production of
those crops.

Research problem:

The research problem is represented in the limited water resources, the low level of efficiency in the use of irrigation, in addition to the decrease in the productive efficiency of some crops due to the poor provision of appropriate agricultural operations, especially the quantities of water and their appropriate quality for irrigation. Using irrigation water, which leads to lower irrigation efficiency.

Research Importance:

A. Applied importance:

This paper, is one of the applied papers in the field of optimal use of water resources, where it studies the utilization of agriculture drainage water, or mixing treated agriculture drainage water with Nile water so that the salinity does not exceed 3000 parts per million(ppm), for using in agriculture, finding non-conventional solutions to increase economic agricultural areas in desert region.

B. Economic importance:

Mint and basil are medicinal and aromatic crops of special economic importance, according to the World Food Organization (FAO)[10], because they are involved in many medicines, beverages and cosmetics. As well as their economic importance, where they have been used recently in the whole world because of the few side effects on the body. So, this paper will study the economic impact of using different types of water to increase the agricultural area in Egypt from these two plants.

Research Objective:

This research aims to study the effect of three types of water on the production of some medicinal and aromatic plants (basil, mint). In addition to the study of some economic indicators of such plants, such as production costs and farm price and net return, in order to reach results which can increase the awareness of producers and thus expand this production through the use of non-conventional sources of water.

Research Assumptions:

The research was based on:

1 .Main hypothesis:

" There are differences between the productivity of some medicinal and aromatic plants according to the type of irrigation water used".

2. Sub-hypothesis:

- There are significant differences between of basil production (green& dry) under irrigation conditions with Nile water, Nile water mixed with treated agriculture drainage water with (ratio of 1: 1 with a salinity of < 3000 ppm) and treated agriculture drainage water.
- There are significant differences between mint production (green& dry) under irrigation conditions with Nile water, Nile water mixed with treated agriculture drainage water with (ratio of 1: 1 with a salinity of < 3000 ppm) and treated agriculture drainage water.

Design experiment:

1.The experiment was conducted to study the effect the different types of irrigation water on the production of both basil and mint plant using a completely randomized design.

2. taken a random sample from each basin (10 plants).

3.The weight of each sample was recorded (green - dry).

4. Using SPSS program, the data recorded from the experiment was analyzed after verifying it is distributed naturally.

Data sources:

The study relied mainly on the published and unpublished secondary data. which are issued by the competent authorities, such as the central agency for public mobilization and statistics, and some data published on some of the united nations site, and research was based on the preliminary data of the practical experiment conducted at station of the Desert Research Center in North Sinai Governorate.

Research Methodology:

To achieve the objectives of the study, an experiment design analysis (Complete Randomize Design) method and some economic analysis methods were used, Such as: descriptive analysis, normality analysis, one way anova, test of homogeneity of variances, least significant difference, return on irrigation water unit, and economic indicators, such as cost structure, total return, net return and The ratio of revenue to total costs.

-Results and discussion:

1. Basil plant:

A. Green basil:

The average weight of green basil sample irrigated with Nile water came first, followed by irrigated mixed water and finally, irrigated with treated agriculture drainage water, with an average weight of 20.894, 20.313 and 10.9770 gm/pot respectively. with standard deviation of 3.72689, 4.42308 and 4.42527 respectively. In addition, standard error about 1.17855,1.39870 and 1.39939 respectively. (Table 1)

B. Dry basil:

The average weight of green basil sample irrigated with Nile water came first, followed by irrigated mixed water and finally, irrigated with treated agriculture drainage water, with an average weight of 6.3810, 6.1720 and 3.3320 gm/pot respectively. with standard deviation of 1.17818, 1.16250 and 1.33630 respectively. In addition, standard error about 0.37257, 0.36761and 0.42257respectively.

Table (1) Descriptive analysis of samples of experiment treatments

		N	Mean	Std. Deviation	Std. Error			N	Mean	Std. Deviation	Std. Error
Green Basil	1	10	10.9770	4.42527	1.39939	Green Mint	1	10	8.2500	1.68894	0.53409
	2	10	20.3130	4.42308	1.39870		2	10	14.1700	1.98327	0.62717
	3	10	20.8940	3.72689	1.17855		3	10	14.6920	2.13261	0.67439
	Total	30	17.3947	6.14989	1.12281		Total	30	12.3707	3.51385	0.64154
Dry Basil	1	10	3.3320	1.33630	0.42257	Dry Mint	1	10	4.6270	0.78626	0.24864
	2	10	6.1720	1.16250	0.36761		2	10	6.9110	0.81346	0.25724
	3	10	6.3810	1.17818	0.37257		3	10	7.1940	0.92921	0.29384
	Total	30	5.2950	1.84527	0.33690		Total	30	6.2440	1.42528	0.26022
	Total										

Source: Calculated by the researcher.

*1 = treated agriculture drainage water

*2= mixed water (Nile water,1: 1 treated agriculture drainage water with salinity < 3000 ppm)

*3= Nile water

2. Mint plant

A. Green Mint

The average weight of green mint sample irrigated with Nile water came first, followed by irrigated mixed water and finally, irrigated with treated agriculture drainage water, with an average weight of 14.6920, 14.1700 and 8.2500 gm/pot respectively. with standard deviation of 2.13261, 1.98327 and 1.68894 respectively. In addition, standard error about 0.67439, 0.62717 and 0.53409 respectively. (Table 1)

B. Dry Mint

The average weight of green mint sample irrigated with Nile water came first, followed by irrigated mixed water and finally, irrigated with treated agriculture drainage water, with an average weight of 7.1940, 6.9110 and 4.6270 gm/pot respectively. with standard deviation of 0.92921, 0.81346 and 0.78626 respectively. In addition, standard error about 0.29384, 0.25724 and 0.24864 respectively. (Table 1)

Analysis of variance :

The results of Table (2) show that there are significant differences between the averages for the irrigation water type at the significant level (0.05). This means that there is a difference in the mean weight of the samples according to the quality of irrigation water used.

Table (2) Variance Analysis of Experimental Process Groups

		Sum of Squares	df	Mean Square	F	Sig.
Green Basil	Between Groups	619.484	2	309.742	17.521	0.000
	Within Groups	477.327	27	17.679		
	Total	1096.812	29			
Dry Basil	Between Groups	58.019	2	29.009	19.232	0.000
	Within Groups	40.727	27	1.508		
	Total	98.746	29			
Green Mint	Between Groups	256.061	2	128.030	33.889	0.000
	Within Groups	102.005	27	3.778		
	Total	358.066	29			
Dry Mint	Between Groups	39.621	2	19.810	27.728	0.000
	Within Groups	19.290	27	0.714		
	Total	58.911	29			

Source: Calculated by the researcher using SPSS.

-LSD test:

It is noticed from Table (3) that there are significant differences between the mean weights according to the irrigation coefficients where the results showed that there are significant differences at the significant level (0.05) for irrigation treatments with Nile and blended water on one side, and between irrigation with treated agriculture drainage water, while there are no significant differences between irrigation treatments with Nile water and blended irrigation water, which shows the validity of the sub-hypotheses.

Table (3) Minimum Difference test (LSD)

	(I) x	(J) x	Mean Difference (I-J)	Sig.		(I) x	(J) x	Mean Difference (I-J)	Sig.
Green Basil	1	2	9.33600	0.000	Green Mint	1	2	5.92000	0.000
		3	9.91700	0.000			3	6.44200	0.000
	2	1	9.33600	0.000		2	1	5.92000	0.000
		3	0.58100	0.760			3	0.52200	0.553
	3	1	9.91700	0.000		3	1	6.44200	0.000
		2	0.58100	0.760			2	0.52200	0.553
Dry Basil	1	2	2.84000	0.000	Dry Mint	1	2	2.28400	0.000
		3	3.04900	0.000			3	2.56700	0.000

	2	1	2.84000	0.000		2	1	2.28400	0.000
		3	0.20900	0.707			3	0.28300	0.461
	3	1	3.04900	0.000		3	1	2.56700	0.000
		2	0.20900	0.707			2	0.28300	0.461

Source: Calculated by the researcher using SPSS.
*1 = treated agriculture drainage water.
*2= mixed water (Nile water,1: 1 treated agriculture drainage water with salinity < 3000 ppm).
*3= Nile water.

Economic study:

Numerous physiological studies [Ahmed&Faisal, 2010], [Zeina & Jamil, 1999], [J.W. Maranville et al, 1993] have proven that cultivation in desert lands and irrigation with salty water increases the quality of plants in terms of taste and smell, in addition to increasing green materials in the plant (chlorophyll) that increases the efficiency of the absorption process The nutrients the plant needs. On the other hand, there is a decrease in plant weight, size, and green leaf area responsible for food preparation to complete growth processes. Also, plant weight is an indication of acre productivity [Maas E V, 1986], [Maas & Hoffman], [Maas & Nieman, 1978], [Raphanus & L.F.M., 1999].

The aim of this study was to analyze the economic effects of increasing the salinity of irrigation water on some medicinal and aromatic plants of high economic importance. From the data presented in Table (1) showed that, productivity decreased when irrigation with mixed water and treated wastewater by about 10%, 50%, respectively, of crop productivity compared to irrigation with Nile water [N. M. Malash et al,2008]. The experiment demonstrated that the use of highly saline water in irrigation reduces water loss from the soil by 10% of the actual needs of the plant, and thus low water needs of the plant under these conditions [Ashour et al, 2009] [Mohamed, 1998]. This leads to a difference in the rate of return per water unit for each type of irrigation water for the two crops.

By studying the cost structure of each plant, (With the stability of elements cost with each type of irrigation water) Assuming the water needs of both basil and mint 2.8, 2.6 M3/ fadden respectively. Table (4) shown that, increasing the average productivity for green basil with Nile water ton/ feddan by 10%, 50% of productivity compared with treated agriculture drainage water + Nile(< 3000 ppm) water and treated agriculture drainage water, To become 2.1,1.89,1.575 ton/ fadden respectively, When the total costs are fixed at 9000 LE/Fadden it becomes the total return 25200, 22869, 19057.5 LE/ fadeen respectively. While the rate of return to costs were 35.7%, 39.35%, 47.23% respectively. Regarding the average prices, we find that they have increased for crops irrigated with high salinity water as well as mixture water, compared to the crop irrigated with fresh water. When calculating the rate of return on water unit, it was found that, the rate of return in the case of irrigation by Nile water increased to be 5785.7 compared with irrigation by mixture water and treated agriculture drainage water to be 5503.6, 4789 LE/ M^3 respectively.

Table (4): most important productive and economic indicators for green basil and green mint Sample research for the agricultural season 2017.

Item	basil			mint		
	1	2	3	1	2	3
Average production Ton/ feddan	1.575	1.89	2.1	1.8	2.16	2.4
total costs LE/ Ton	9000	9000	9000	8000	8000	8000
Total return LE/ Ton	19057.5	22869	25200	18180	21816	24000
Net return LE/ Ton	10057.5	13869	16200	10180	13816	16000
R/C ratio	47.23%	39.35%	35.7%	78.5%	36.67%	33.3%
Average price LE/ Ton	12100	12100	12000	10100	10100	10000
Water quantity M^3/ Fadden	2.1	2.52	2.8	1.95	2.34	2.6
Net water unit return LE/ M^3	4789	5503.6	5785.7	5220.5	5904.3	6153.8

- Net water unit return (LE/M^3) = Net return (LE/ Ton) ÷ Water quantity (M^3/ Fadden).
- Source: collected and computed by researcher.

For the mint plant, it was found that, increasing the average productivity for green mint with Nile water ton/ feddan by 10%, 50% of productivity compared with treated agriculture drainage water + Nile(< 3000 ppm) water and treated agriculture drainage water, To become 2.4, 2.16, and 1.8 ton/ fadden respectively, When the total costs are fixed at 8000 LE/Fadden it becomes the total return 24000, 21816, 18180 LE/ fadeen respectively. While the rate of return to costs were 33.3%, 36.67% and 78.5% respectively. Regarding the average prices, we find that they have increased for crops irrigated with high salinity water as well as mixture water, compared to the crop irrigated with fresh water. When calculating the rate of return on water unit, it was found that, the rate of return in the case of irrigation by Nile water increased to be 6153.8 compared with irrigation by mixture water and treated agriculture drainage water to be 5904.3, 5220.5 LE/ M^3 respectively.

Summary

In this research, a field study was carried out to determine the effect of three levels of water salinity on the cultivation of basil and mint plants in desert lands because of their great economic importance in the local and international market. The results showed that the average production of the green basil sample irrigated with blended irrigation water was 97.2% compared to the average production of the green basil sample irrigated with Nile water. The average of the green mint sample irrigated with mixed irrigation water was 96.4% compared to the average weight of the green mint sample irrigated with Nile water.

The results of the research paper concluded that by using mixed water (Nile water,1: 1 treated agriculture drainage water with salinity < 3000 ppm), works to provide quantities of irrigation water that can be used in horizontal expansion in the cultivation of some crops that are somewhat tolerant to salinity such as medicinal and aromatic plants, such as basil and mint plants, where Such plants give great economic returns in a short time also have a relative importance in world exports as demand for them from most countries of the world rises. As well as providing irrigation water and increasing the rate of return on water unit, in addition to the high total return which

approximates the total return of plants irrigated with Nile water, and thus the Nile water can be provided for domestic and industrial uses.

References:

1. Central agency for public mobilization and statistics, Irrigation and Water Resources Bulletin, separate issues.
2. Mamdouh Fathi Abdel Sabour - **The use of sea water in agriculture and the production of saline-loving plants Nuclear Research Center** - Egyptian Atomic Energy Authority, Assiut Journal for Environmental Studies - No. 25 July 2003.
3. Yasser Hamdy Abdellah Ali, **Analytical Study of the Egyptian Foreign Trade of the most important medical and aromatic crops**, Ph.D. Thesis, Department of Agricultural Economics, Faculty of Agriculture, Minia University, 2010.
4. Agricultural Extension Department - General Directorate for Agriculture and Fisheries Affairs.
5. Ministry of Agriculture and Land Reclamation, Central Department of Agricultural Economics, Bulletin of Agricultural Statistics.
6. Bouwer H Irrigation and global water outlook. Agric Water Manage 25:221–231,1994.
7. Ragab R, **Constraints and applicability of irrigation scheduling under limited water resources**, variable rainfall and saline conditions, 1997.
8. Grattan SR, Shennan C, May D, Roberts B, Borin M, Sattin, M **Utilizing saline drainage water to supplement irrigation water requirements of tomato in a rotation with cotton**. In: Proceedings of the 3rd congress of the European Society for Agronomy, Padova University, Abano-Padova, Italy, 18–22 September 1994, pp 802–803.
9. Oudhia, P. **Traditional and medicinal knowledge about pudina(Mentha sp. Family . Labiatae) in chattisgarh** , India Agric ., 8 : 102 – 109,(2003)
10. Food and Agriculture Organization of the United Nations - FAO.
11. Ahmed Abdul-Redha, Faisal Al-Mazloum **- Effect of saline stress on some biochemical traits of peppermint and menthol outside the body**, Faculty of Science, University of Kufa, Iraq, 2010.
12. Zeina Abdel Moneim, Jamil al-Mufti, **the problem of salinity and salt land &Plant Physiology under the Effect of Drought and Saline Stress**, King Saud University, Riyadh, 1999.
13. E. V. Maas U.S. Salinity Laboratory, USDA-ARS Riverside, CA, J.W. Maranville, B.V.BaIigar, R.R. Duncan, J.M. Yohe. (eds.) **INTSORMIL, Testing Crops for Salinity Tolerance, proc. workshop on Adaptation of Plants to Soil Stresses.** p. 234-247. Pub. No. 94-2, Univ of Ne, Lincoln, NE, August l-4-1993.
14. Maas E V, **Salt tolerance of plants**. Appl. Agr. Res. 1, 12–26. 1986
15. Maas E V and Hoffman G J, **Crop salt tolerance - Current assessment.** J. Irr. Drain. Div., Proc. Am. Soc. Civ. Eng. 103(IR2), 115–134. 1977
16. Maas E V and Nieman R H, **Physiology of plant tolerance to salinity**. In Crop tolerance to suboptimal land conditions. Ed. G. 1978
17. Raphanus sativus L., L.F.M. Marcelis **Plant and Soil Effect of salinity on growth, water use and nutrient use in radish**, Research Institute for

Agrobiology and Soil Fertility (AB-DLO), P.O. Box. 14, 6700 AA Wageningen, The Netherlands,57–64, 57,1999.

18. N. M. Malash,T. J. Flowers, R. Ragab, **Effect of irrigation methods, management and salinity of irrigation water on tomato yield, soil moisture and salinity distribution,** Irrigation Science, May 2008, Volume 26, Issue 4, pp 313–323.
19. Ashour, Samir Kamel - Salem, Samia Aboul Fotouh **Statistical presentation and analysis using SPSSWIN, Part III:** Experimental Design and Analysis, Institute of Statistical Studies and Research, Cairo University, pp. 124- 161, 2009.
20. Mohamed Reda Othman, **Rules and principles of plant breeding**, Tishreen University Publications, page 358,1998.

Bayesian Identification Methods for Autoregressive Models: A Comparison Study

Ayman A. Amin*

Abstract

In this paper we first review the existing Bayesian identification methods for the autoregressive (AR) models that can be classified into three groups: (1) Testing based identification that assumes the AR model order is an unknown constant with a known maximum, (2) Posterior mass function based identification that assumes the AR model order is a random variable with a known maximum, (3) Stochastic search variable selection (SSVS) based identification that introduces a latent variable for each coefficient of the AR model. Second, we present an extensive simulation-based comparison of these Bayesian identification methods applied to AR models, considering different factors, such as the time series size, actual model order and model coefficients values, aiming to cover various realistic situations. In general, simulation results show that the testing based identification method is more accurate with small time series sizes than posterior mass function and SSVS based identification methods that provide higher accuracy with moderate and large time series sizes, respectively.

Key Words: Time series model identification, normal-gamma prior, Posterior mass function, Stochastic search variable selection.

1 Introduction

Time series models – especially autoregressive (AR) models - are widely used to fit and forecast time series data in many fields such as economics and industrial engineering. Time series modelling starts with the identification of the model order, followed by the model estimation, diagnostics check and model forecasting. Therefore, the model identification step is important since all other steps depend on its accuracy (Box et al., 2015). The model identification means the model order of the underlying time series is unknown and needs to be specified or estimated based on the given time series data and its assumed probability distribution (Amin, 2019).

Different Bayesian identification methods are proposed in literature in order to identify the best AR model order for the uderlying time series data. Based on our literature review of the

* Assistant Professor of Statistics, Department of Statistics, Mathematics, and Insurance, Faculty of Commerce, Menoufia University, Egypt. Email: ayman.a.amin@gmail.com.

Bayesian time series analysis, we can classify these methods into three main methods. The first identification method is a testing based identification that assumes the model order is an unknown constant with a known maximum, and the best order can be estimated based on a sequence of t-test of significance (Broemeling and Shaarawy, 1987). This Bayesian identification method is applied to different time series models including autoregressive models (Broemeling and Shaarawy, 1988; Daif et al., 2003), autoregressive moving average models (Ali, 2003), moving average models (Amin, 2018), and seasonal moving average models (El-Souda, 2008).

The second identification method is a posterior mass function based identification that assumes the model order is a random variable with a known maximum, and its posterior mass function can be derived to select the order as a value with the highest posterior probability. Following this idea, Diaz and Farah (1981) proposed a Bayesian method to identify the order of autoregressive models. Their work has been extended by researchers to different time series models, which include autoregressive moving average models (Fan and Yao, 2009), seasonal autoregressive models (Shaarawy and Ali, 2003), multivariate autoregressive models (Shaarawy and Ali, 2008), and double seasonal autoregressive models (Amin, 2017b and Amin, 2019).

The third identification method is stochastic search variable selection (SSVS) based identification method that introduces a latent variable for each coefficient of the AR model, and uses the Gibbs sampler to generate a sample from the conditional posterior of the latent variable, which guarantees that the latent variable with the highest posterior probability will be sampled and accordingly the best AR model subset will be selected as a value with the highest frequency in the generated marckov chain. SSVS procedure was first proposed by George and McCulloch (1993) for linear regression models aiming to select the best subset model, and then it is extended by Chen (1999) for AR models, So et al. (2006) for AR models with exogenous variables, and Chen et al. (2011) for ARMA models to identify the best time series model.

These Bayesian identification methods were proposed by different researchers; however, there is no study that evaluates and compares their identification accuracy. Therefore, in this paper we first review the Bayesian background of these identification methods, and then conduct an extensive simulation study aiming to compare their accuracy with the application to AR models, considering different factors such as the time series size, actual model order and model coefficients values in order to present a fair comparison and cover various realistic situations.

The remainder of this paper is organized as follows. In Section 2 we present the background of the AR time series models and related Bayesian concepts. In Section 3 we summarize the background of the existing Bayesian identification methods for AR models. In Section 4 we present the simulation study to evaluate and compare the accuracy of these Bayesian identification methods. Finally, we give the conclusions in Section 5.

2 Autoregressive Models and Related Bayesian Concepts

Time series $\{y_t\}$ is modeled by an autoregressive model of order p, AR(p), if it can be written as (Box et al., 2015):

$$\phi_p(B)y_t = \varepsilon_t, \qquad (1)$$

where B is a backshift operator defined as $B^d x_t = x_{t-d}$, $\varepsilon_t's$ are independent and normally distributed errors, i.e. $\varepsilon_t \sim N(0, \sigma^2)$, and $\phi_p(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$ is the autoregressive polynomial with order p. We can simplify the AR model (1) to be written as:

$$y = X\beta + \varepsilon, \tag{2}$$

where $y = (y_1, y_2, \cdots, y_n)^T$, X is an $n \times p$ matrix with the t^{th} row $X_t = (y_{t-1}, \ldots, y_{t-p})$, $\beta = (\phi_1, \phi_2, \ldots, \phi_p)^T$ is the AR coefficients, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$.

It is worth noting that the design matrix X becomes a function of p when the AR model order is unknown.

Bayesian analysis of time series models is based on Bayes' theorem that combines the prior distribution of the model parameters with the likelihood function of observed sample to get the posterior distribution.

Regarding the prior specification, we consider the normal-gamma prior for the model parameters β and σ^2. Thus, suppose $\beta \sim N_p(\mu_\beta, \sigma^2 \Sigma_\beta)$ and $\sigma^2 \sim IG(\frac{\nu}{2}, \frac{\lambda}{2})$, the normal-gamma prior distribution of β and σ^2 is given by:

$$\zeta(\beta, \sigma^2) \propto (\sigma^2)^{-\left(\frac{\nu+p}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma^2}\left[\lambda + (\beta - \mu_\beta)^T \Sigma_\beta^{-1} (\beta - \mu_\beta)\right]\right\}, \tag{3}$$

where $\mu_\beta, \Sigma_\beta, \nu$ and λ are hyperparameters need to be estimated.

The likelihood function of the AR model (2) can be obtained by employing a straightforward random variable transformation from ε to y, and written as

$$\begin{aligned} L(\beta, \sigma^2, p|y) &\propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \varepsilon^T \varepsilon\right\}, \\ &\propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)\right\}, \end{aligned} \tag{4}$$

Multiplying the likelihood function (4) by the normal-gamma prior (3) results in the joint posterior of the model parameters β and σ^2 as:

$$\begin{aligned} \zeta(\beta, \sigma^2, p|y) \propto (\sigma^2)^{-\left(\frac{n+\nu+p}{2}+1\right)} \exp\Big\{-\frac{1}{2\sigma^2}\Big[\lambda + (\beta - \mu_\beta)^T \Sigma_\beta^{-1} (\beta - \mu_\beta) + \\ (y - X\beta)^T (y - X\beta)]\Big\}. \end{aligned} \tag{5}$$

3 Existing Bayesian Identification Methods for AR Models

Different Bayesian identification methods were proposed in the literature of Bayesian time series analysis to identify the best order of AR time series models. These Bayesian identification methods canbe classified into three groups: testing based identification, posterior mass function based identification, and SSVS based identification, which are discussed in the following subsections.

3.1 Testing Based Identification of AR Models

The testing based identification method for AR models is proposed by Broemeling and Shaarawy (1987) and can be summarized in two main steps. First, the model order p is assumed to

be an unknown constant with a known maximum and accordingly the posterior distribution of model coefficients is derived. Second, a sequence of t-test of significance is executed based on the marginal posterior of the model coefficients to specify the best value of the model order p.

Assuming the model order p is an unknown constant with a known maximum k, from the joint posterior of β and σ^2 in eqn (5) we can easily show that the marginal posterior of the model coefficients β is a multivariate t distribution with degrees of freedom $v_n = (n + \nu)$ and location vector and dispersion matrix are respectively:

$$\mu_n = A_n^{-1}B_n \qquad and \qquad V_n = \frac{1}{v_n - 2}A_n^{-1}C_n, \tag{6}$$

where,

$A_n^{-1} = (X^TX + \Sigma_\beta^{-1})^{-1}$, $B_n = (X^Ty + \Sigma_\beta^{-1}\mu_\beta)$, and $C_n = \left[y^Ty + \lambda + \mu_\beta^T\Sigma_\beta^{-1}\mu_\beta - B_n^TA_n^{-1}B_n\right]$.

The main property of the multivariate t distribution of a vector is that any single component of this vector has a univariate t distribution and the conditional distribution of any component given any other component has also a univariate t distribution. Using this marginal posterior of the parameters vector β and based on the property of the multivariate t distribution, we can do a backward elimination procedure to identify the best value of the model order p as follows:

1. Test $H_0: \phi_k = 0$ against $H_1: \phi_k \neq 0$ using the marginal posterior distribution of ϕ_k which is a univariate t distribution.
2. If H_0 is not rejected, test $H_0: \phi_{k-1} = 0$ against $H_1: \phi_{k-1} \neq 0$ using the conditional posterior distribution of ϕ_{k-1} given $\phi_k = 0$ which is also a univariate t distribution.
3. A sequence of t-test of significance is executed until the hypothesis $\phi_{p_0} = 0$ is rejected where $0 < p_0 \leq k$, which means the identified value for p is p_0.

3.2 Posterior Mass Function Based Identification of AR Models

The posterior mass function based identification assumes the AR model order p is a random variable with a known maximum k. In this case, the prior information about p can be represented in terms of a prior mass function $\zeta(\mathrm{p})$ that can have different forms such as an uniform, i.e. $\zeta(\mathrm{p}) = 1/\mathrm{k}$, or a geometric, i.e. $\zeta(\mathrm{p}) = 0.5^{\mathrm{p}}\ \forall \mathrm{p} = 1, \ldots, \mathrm{k}$. Therefore, the marginal posterior mass function of the model order can be derived and then the best model order p can be selected as a value with a maximum posterior probability (Amin, 2017d; Amin, 2018a).

Using the joint posterior of β and σ^2 in eqn (5) and the prior mass function $\zeta(\mathrm{p})$, the joint posterior of the model parameters β, σ^2 and p can be given as:

$$\zeta(\beta, \sigma^2, p|y) \propto \zeta(p)(\sigma^2)^{-\left(\frac{n+\nu+p}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma^2}\left[\lambda + \left(\beta - \mu_\beta\right)^T\Sigma_\beta^{-1}\left(\beta - \mu_\beta\right) + (y - X\beta)^T(y - X\beta)\right]\right\}. \tag{7}$$

Accordingly, by integrating out the parameters β and σ^2 in eqn (7) the marginal posterior mass function of the model order p can be derived and written as:

$$\zeta(p|y) \propto \zeta(p)\left[\frac{\left|\Sigma_\beta^{-1}\right|}{|A_n|}\right]^{\frac{1}{2}}\left[y^Ty + \lambda + \mu_\beta^T\Sigma_\beta^{-1}\mu_\beta - B_n^TA_n^{-1}B_n\right]^{-\frac{n+\nu}{2}} \quad \forall p = 1, \ldots, k. \tag{8}$$

Where $A_n = (X^T X + \Sigma_\beta^{-1})$ and $B_n = (X^T y + \Sigma_\beta^{-1}\mu_\beta)$. From this marginal posterior mass function, we select a value of p with the highest posterior probability as the best AR model order.

3.3 SSVS Based Identification of AR Models

Stochastic search variable selection (SSVS) procedure was first proposed by George and McCulloch (1993) for linear regression models aiming to select the best subset model, and then it is extended by Chen (1999) for AR models and So et al. (2006) for AR models with exogenous variables to identify the best AR model.

The main idea of the SSVS based identification method for AR model is to introduce a latent binary variable for each coefficient of AR model (2), i.e. $\delta_i \; \forall i = 1, \ldots, k$, where $\delta_i = 1$ or 0 depending on whether the ith lag (corresponding to ϕ_i) is selected or not. Accordingly, the prior information on a single AR model coefficient ϕ_i can be presented by the normal mixture distribution given as:

$$\phi_i|\delta_i \sim (1 - \delta_i)N(0, \tau_i^2) + \delta_i N(0, c_i^2\tau_i^2), \tag{9}$$

and

$$p(\delta_i = 1) = 1 - p(\delta_i = 0) = P_i \tag{10}$$

Therefore, the prior distribution of $\beta|\boldsymbol{\delta}$ can be specified as a multivariate normal and given as:

$$\beta|\boldsymbol{\delta} \sim N(\mathbf{0}, D_{\boldsymbol{\delta}} V D_{\boldsymbol{\delta}}), \tag{11}$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)$, $D_{\boldsymbol{\delta}}$ is the diagonal matrix, i.e. $D_{\boldsymbol{\delta}} = diag[a_1\tau_1, \ldots, a_k\tau_k]$, with $a_i = 1$ if $\delta_i = 0$ and $a_i = c_i$ if $\delta_i = 1$, and V is the prior correlation matrix, which can be assumed to be identity matrix. $D_{\boldsymbol{\delta}}$ is specified in such a way to be a scaling of the prior covariance matrix. We set $\tau_1, \ldots, \tau_k$ to be small and $c_1, \ldots, c_k$ to be large and greater than 1, so that the coefficients $\phi_i's$ for which $\delta_i = 0$ will tend to be clustered around zero, whereas for which $\delta_i = 1$ will tend to be dispersed and away from zero. For more details about the choices of these constants see George and McCulloch (1993).

For prior specification of σ^2, we employ the inverse gamma prior as:

$$\sigma^2|\boldsymbol{\delta} \sim IG(\tfrac{\nu}{2}, \tfrac{\lambda}{2}), \tag{12}$$

We assume δ_i's are a priori independent with a marginal distribution given in eqn (10), and therefore the joint prior distribution of $\boldsymbol{\delta}$ can be written as:

$$\zeta(\boldsymbol{\delta}) = \prod_{i=1}^{k} P_i^{\delta_i}(1 - P_i)^{(1-\delta_i)} \tag{13}$$

As a special case, $\zeta(\boldsymbol{\delta}) = 2^{-k}$, which is an uniform prior of $\boldsymbol{\delta}$ where each lag has same probability to be selected, i.e. $P_i = 1/2$.

Multiplying the likelihood function (4) by the prior distributions (11) - (13) results in the joint posterior that can be written as:

$$\zeta(\beta, \sigma^2, \boldsymbol{\delta}|y) \propto \prod_{i=1}^{k} P_i^{\delta_i}(1 - P_i)^{(1-\delta_i)} (\sigma^2)^{-\left(\frac{n+\nu}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma^2}[\lambda + \sigma^2\beta^T(D_{\boldsymbol{\delta}} V D_{\boldsymbol{\delta}})\beta + (y - X\beta)^T(y - X\beta)]\right\}. \tag{14}$$

In order to avoid computing the posterior probabilities of all 2^k subsets, SSVS uses the

Gibbs sampler to generate a sample from the conditional posterior of $\boldsymbol{\delta}$, which guarantees that the $\boldsymbol{\delta}$ with the highest posterior probability will be sampled and the best subset will be selected (George and McCulloch, 1993). However, in order to apply the Gibbs sampling algorithm the full conditional posteriors of the model parameters β, σ^2 and $\boldsymbol{\delta}$ have to be derived.

From the joint posterior (14) and using some Bayesian standards, it is easy to show that:

- The conditional posterior of β given σ^2, $\boldsymbol{\delta}$ and $\boldsymbol{y}$ is a multivariate normal with a mean vector μ_β^* and a covariance matrix V_β^* given as:
  $\mu_\beta^* = [(\sigma^{-2}H^TH + (D_{\boldsymbol{\delta}}VD_{\boldsymbol{\delta}})^{-1})^{-1}(\sigma^{-2}H^Ty)]$ and $V_\beta^* = (\sigma^{-2}H^TH + (D_{\boldsymbol{\delta}}VD_{\boldsymbol{\delta}})^{-1})^{-1}$
  Where H is an $n \times k$ matrix with the ti^{th} element $H_{ti} = y_{t-i}$.
- The conditional posterior of σ^2 given β, $\boldsymbol{\delta}$ and $\boldsymbol{y}$ is an inverse gamma with parameters $\left(\frac{n+\nu}{2}, \frac{\lambda^*}{2}\right)$, where $\lambda^* = \nu\lambda + (y - X\beta)^T(y - X\beta)$.
- The conditional posterior of δ_i given β, σ^2, other variables in $\boldsymbol{\delta}$ – referred as $\boldsymbol{\delta}_{(-i)}$ – and $\boldsymbol{y}$ is a Bernoulli with probability
  $$P(\delta_i = 1 | \beta, \sigma^2, \boldsymbol{\delta}_{(-i)}, y) = \frac{a}{a+b}$$
  Where $a = P_i \times \zeta(\beta|\sigma^2, \boldsymbol{\delta}_{(-i)}, y, \delta_i = 1)$ and $b = (1 - P_i) \times \zeta(\beta|\sigma^2, \boldsymbol{\delta}_{(-i)}, y, \delta_i = 0)$

4 Simulation-based Comparison Study

In this section we present a simulation-based comaprison study to evaluate the accuracy of the existing Bayesian identification methods for AR models. In order to present fair comparison and to cover various realistic situations, we generate several simulated time series data with different sample size, different model orders, and different values of the model coefficients. In particular, we generate 1,000 time series of size n from 50 to 200 (with an increment of 50) from AR models with order one, where ϕ = 0.3, 0.5, and 0.8, and with order two, where (ϕ_1, ϕ_2) = (0.2, 0.6), (0.5, 0.4) and (0.9, 0.4), and with order three, where (ϕ_1, ϕ_2, ϕ_3) = (0.4, -0.5, 0.6), (0.5, -0.4, 0.6), and (-0.7, 0.5, 0.8).

Once the time series datasets are generated from these several AR models, the three Bayesian identification methods presented in Section 3 are applied by assuming the normal-gamma prior for the model parameters β and σ^2, as given in eqn (3), and employing the uniform prior for the model order p given as $\zeta(p) = \frac{1}{k}, \forall p = 1, \ldots, k$, and considering $k = 4$. All required hyperparameters are estimated using empirical Bayesian approach (Amin, 2017a). In particular, to apply the testing based identification method we execute a sequence of t-test of significance using the marginal posterior of the model coefficients β and based on these tests results we specify the best value of the order p. In addition, to apply the posterior mass function based identification we compute the posterior mass function of the AR model order p, $\zeta(p|y)$, and then we identify the best order as a value with the highest posterior probability. Moreover, to apply the SSVS based identification method we run the Gibbs sampler with the conditional posteriors of β, σ^2 and latent vector $\boldsymbol{\delta}$, which are derived at the end of subsection 3.3, and the best AR model subset will be selected according to the value with the highest frequency in the generated marckov chain of $\boldsymbol{\delta}$.

For all simulated time series, we compute the percentage of correctly identified models by comparing the identified order with the true value of p used to gesnerate the time series. Results

of the simulation study are presented in Tables (1) - (3).

Table 1: Percentage of correctly identified models for AR(1).

n	Identification Method		
	Testing	PMF[1]	SSVS
$\phi_1 = 0.3$			
50	88.5	74.6	37.8
100	87.0	85.7	65.0
150	86.1	89.5	80.6
200	86.6	91.1	88.7
$\phi_1 = 0.5$			
50	87.8	74.8	76.0
100	85.3	85.9	89.6
150	85.8	89.8	92.2
200	85.8	91.9	93.7
$\phi_1 = 0.8$			
50	84.6	76.0	85.9
100	81.5	87.3	91.6
150	80.5	90.6	93.9
200	80.4	92.3	95.5

[1] Posterior mass function.

From the simulation results we can observe some remarks. First, the larger time series size the higher percentage of correctly identified models is obtained by the three Bayesian identification methods, since the observed time series data provides enough information about the correct AR models order.

Table 2: Percentage of correctly identified models for AR(2).

n	Identification Method		
	Testing	PMF	SSVS
$\phi_1 = 0.9$ and $\phi_2 = -0.3$			
50	43.2	61.1	43.2
100	67.6	79.3	69.1
150	74.8	87.0	81.9
200	79.1	91.3	91.5
$\phi_1 = 0.3$ and $\phi_2 = 0.5$			
50	80.8	76.2	32.9
100	89.9	86.2	73.1
150	90.3	90.1	90.8
200	91.1	91.7	95.2
$\phi_1 = 0.5$ and $\phi_2 = 0.4$			
50	61.2	67.7	51.7
100	86.1	85.6	87.3

150	89.0	89.6	95.0
200	89.0	92.4	96.9

Table 3: Percentage of correctly identified models for AR(3).

n	Identification Method		
	Testing	**PMF**	**SSVS**
$\phi_1 = 0.4,\ \phi_2 = -0.5$ and $\phi_3 = 0.6$			
50	93.7	84.0	52.8
100	96.1	89.8	90.3
150	95.3	91.9	95.4
200	95.6	93.6	96.7
$\phi_1 = 0.5,\ \phi_2 = -0.4$ and $\phi_3 = 0.6$			
50	94.5	83.2	60.3
100	95.8	90.6	92.9
150	95.3	92.4	96.8
200	96.2	94.2	97.6
$\phi_1 = -0.7,\ \phi_2 = 0.5$ and $\phi_3 = 0.8$			
50	97.7	86.6	58.6
100	96.3	89.8	88.2
150	96.5	92.9	94.0
200	96.6	94.5	95.7

Second, for the simple AR models that has only one coeffiecient, i.e. AR(1) in Table (1), a high percentage of correctly identified models is obtained. Similarly, for the AR models with higher order close to the assumed maximum value $k = 4$, i.e. AR(3) in Table (3), the three Bayesian identification methods provide a high percentage of correctly identified models because only a few choices of model orders are available to select from. However, for the complicated AR model with small time series size, i.e. AR(2) in Table (2), a low percentage of correctly identified models is obtained.

Third, the testing based identification method is more accurate with small time series sizes than posterior mass function and SSVS based identification methods that provide higher accuracy with moderate and large time series sizes, respectively.

6 Conclusion

In this paper we first reviewed and summarized the existing Bayesian identification methods for the autoregressive (AR) models that can be classified into three methods: testing based identification, posterior mass function based identification and Stochastic search variable selection (SSVS) based identification. Second, we used a large number of Monte Carlo simulations to evaluate and compare the accuracy of these Bayesian identification methods applied to AR models, and we considered the time series size, actual model order and model

coefficients values to present a fair comparison and cover various realistic situations. In general, simulation results showed that the testing based identification method is more accurate with small time series sizes than posterior mass function and SSVS based identification methods that provide higher accuracy with moderate and large time series sizes, respectively. Future work may be an extension to multivariate autoregressive models.

References

Ali, S. S. (2003). Bayesian Identification of ARMA Models. Unpublished Ph.D. Dissertation, Department of statistics, Faculty of Economics and Political Science, Cairo University, Egypt.

Amin, A. (2017a). Sensitivity to Prior Specification in Bayesian Identification of Autoregressive Time Series Models. *Pakistan Journal of Statistics and Operation Research,* 13(4), 699-713.

Amin, A. (2017b). Identification of Double Seasonal Autoregressive Models: A Bayesian Approach. In Proceedings of the 52nd Annual International Conference of Statistics, Computer Science and Operations Research. ISSR, Cairo University, Egypt.

Amin, A. (2018). A New Look at Bayesian Identification of Moving Average Models. In Proceedings of the 53rd Annual International Conference of Statistics, Computer Science and Operations Research. ISSR, Cairo University, Egypt.

Amin, A. (2019). Bayesian Identification of Double Seasonal Autoregressive Models. *Communications in Statistics: Simulation and Computation*, 48 (8), 2501-2511.

Box, G., Jenkins, G., Reinsel, G., and Ljung, G. (2015). Time series Analysis, Forecasting and control. Fifth Edition, John Wiley & Sons.

Broemeling, L. and Shaarawy, S. (1987). Bayesian Identification of Time Series. In Proceedings of the 22nd Annual Conference in Statistics , Computer Science and Operation Research, Institute of Statistical Studies and Research, Cairo, Egypt, Vol.1,pp.146-159.

Broemeling, L. and Shaarawy, S. (1988). Time Series: A Bayesian Analysis in Time Domain. Studies in Bayesian Analysis of Time Series and Dynamic Models, edited by J. Spall, Marcel Dekker Inc, New York, pp 1-21.

Chen, C. W. (1999). Subset selection of autoregressive time series models. Journal of Forecasting, 18(7), 505-516.

Chen, C. W., Liu, F. C., and Gerlach, R. (2011). Bayesian subset selection for threshold autoregressive moving-average models. Computational Statistics, 26(1), 1-30.

Dias, J. and Farah, J. L. (1981). Bayesian Identification of Autoregressive Process. Presented at the 22nd NBER-NSC Seminar on Bayesian Inference in Econometrics.

Daif, A., Soliman, E. and Ali, S.(2003). On Direct and Indirect Bayesian Identification of Autoregressive Models. In Proceedings of the 15 th Annual Conference in Statistics and Computer Modeling in Human and Social Sciences, Faculty of Economics and Political Science , Cairo University, Egypt.

El-Souda, R.M. (2008). Bayesian Identification for Seasonal Time Series models. Unpublished PhD Dissertation, Department of Statistics, Faculty of Economics and Political Science, Cairo

University, Egypt.

Fan, C. and Yao, S. (2009). Bayesian approach for arma process and its application. International Business Research, 1(4):49–55.

George, E. I., and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423), 881-889.

Shaarawy, S. and Ali, S.(2003). Bayesian Identification of Seasonal Autoregression Models. Communications in Statistics – Theory and Methods, Vol.32, Issue 5, pp 1067-1084.

Shaarawy, S. and Ali, S. (2008). Bayesian Identification of Multivariate Autoregressive Models. Communications in Statistics –Theory and Methods, Vol. 37, Issue 5 , pp 791-802.

Shaarawy, S. M., Soliman, E. A. and Ali, S. S. (2007): Bayesian Identification of Moving Average Models, Communications in Statistics - Theory and Methods, 36 (12): 2301-2312.

So, M. K., Chen, C. W., and Liu, F. C. (2006). Best subset selection of autoregressive models with exogenous variables and generalized autoregressive conditional heteroscedasticity errors. Journal of the Royal Statistical Society: Series C (Applied Statistics), 55(2), 201-224.

Robust Generalized Minimum Variance Control (GMVC): An Application on Economic Time Series

Zidan,A.I.

Abstract:

The main objective of multivariate time series analysis is to obtain a suitable model which adequate to represented the data to use for purposes of forecasting and control. The optimal control means that optimal way to control a dynamic system, by achieving the control design with attainment minimum variance for the model intended to keep the output at the reference value. In control theory, this approach is called minimum variance control. In this paper, we manipulated with single-input single-output (SISO) case which applied on the advertising-sales Lydia Pinkham company represented as an extended autoregressive moving average (EARMA) model. Design and drive control by generalized minimum variance control (GMVC) contributed to reducing the output system variance as a robust way to solve some practical aspects of control design.

Keywords:

Extended Autoregressive Moving Average (EARMA) model, Generalized Minimum Variance Control (GMVC), Minimum Variance Control (MVC), Minimum Variance Incremental Control (MVIC), Multivariate time series analysis.

1-Introduction

Control theory methods are used to find the optimal set of policies over time for a deterministic or stochastic system. Since a large number of economic problems are naturally described as dynamic systems that can be influenced by policies in an attempt to improve their performance. Stochastic control theory is important to find the mathematical design to control the stochastic elements which are common in economics in equations errors, unknown parameters, and measurement errors, so the methods of stochastic and adaptive control essential to applications in economic systems.

The goal of control effort is to keep the output at the target value. However, if the system has the noise or disturbance, it is seldom possible to manipulate the output exactly at the target value. The best action can do is to keep the deviations or error from target values as small as possible. Therefore, the optimal control is defined as a way in which the adjustment inputs values that yield minimum variance or a minimum mean squared error of the output, where the reference is either zero or constant. But advertising- sales system varies at the time, so we use a generalized minimum variance method.

GMVC control is a robust control for controlling the advertising-sales system. This controller is a good dynamic relationship between advertising and sales. The real application shows that the effective role of increasing the desired planning reference on the system control correction action. The sales desired increase as planning for control until the variance of deviation between desired sales and actual sales is minimum variance than the system before uncontrolled. After application of some minimum variance control methods such as minimum variance control strategy (MVCS) recommended by Pandit (1983) and incremental control (IC) on advertising-sales system, the results shows that the single-input single-output (SISO) mechanism with closed loop design was used to control the parameters of the system contributed to the system's stability and the results show that the GMVC is robust for the treatment of the economic system with varied target and solve the practical problem of the non-minimum phase with a reduction in the variance occurrence. The time series control application in the current paper is a novel way of control based mathematical model equivalent to the control in the engineering filed.

2-Modeling System

A primary requirement for any control application is a model that describes the general relationship between economic instruments (inputs, controls) and endogenous variables (outputs). Dynamic data system (DDS) approach is used to modeling seasonal advertising-sales Lydia Pinkham monthly data as nonstationary time series for five years from January 1954-December 1958 in both cases of modeling of univariate and bivariate models. To design extended autoregressive moving average EARMA model as simplifying vector autoregressive moving average VARMA model from a nonstationary time series with deterministic trend and seasonality. It is necessary to transform each univariate advertising-sales time series to stationary before construct VARMA models as recommended in literature William W. S. Wei. (2019), Box et al.(2015), Rufino, (2008), Wei et al., (2006). Todorov and Jordan (2002), Feichtinger et al. (1994), Bhattacharyya, (1982).

Pandit in 1983 recommended decomposing a nonstationary time series into two parts deterministic and stochastic to obtain the companied model for stationary time series used to construct the EARMA model. After applying (DDS) modeling approach in the univariate advertising time series denoted by x_{1t} representing the input of the system and the sales time series denoted by x_{2t} representing the output of the system, the system equations of the adequate model as (SISO) design from the stationary advertising-sales time series described as EARMA (n,n,n-1) model is given by (1,1,0) for advertising and (4,4,3) for sales given by:

$$\left.\begin{aligned} x_{1t} &= -0.132x_{1t-1} + 0.136x_{2t-1} + a_{1t} \\ x_{2t} &= 0.165x_{1-1t} + 0.321x_{1t-2}\ 0.062x_{1t-3} - 0.254x_{1t-4} \\ &\quad -0.10x_{2t-1} + 0.274x_{2t-2} + 0.058x_{2t-3} - 0.133x_{2t-4} \\ &\quad +0.22\,a_{2t-1} - 0.40a_{2t-2} + 0.7a2t\text{-}3 + a_{2t} \end{aligned}\right\} \quad (1)$$

Where: $\gamma_{11} = 21337$ and $\gamma_{22} = 7084.4$ are the error variance of x_{1t} and x_{2t} respectively and the output variance $\sigma^2 = 14595$

3-Minimum Variance Optimal Control

In the engineering filed, control theory is concerned with achieving desired system behavior through the manipulation of the system inputs. The nature of the economic process leads naturally to its description in stochastic rather than in the deterministic. We can apply control in both the economic and engineer field Athans and Falb(2013), Kirk(2012), Lewis et al.(2012), Sethi and Thompson(2006), Hanssens et al.(2003), Seierstad and Sydsaeter(1986), Holly and Zarrop(1979).

The basic effort of control is to obtain the control equation to keep the output at the reference value by minimum mean square error. If we equating the forecasts by zero, the forecast error become equivalents to the system output. The optimal control means that optimal way to control a dynamic system, by achieving the control equation with attainment minimum variance intended to keep the output at the reference value. In control theory, this approach is called minimum variance control.

In 1983, Pandit and Wu designed a minimum mean squared error control strategy based on the minimum variance control approach. The minimum mean squared error control strategy is to adjust the input advertising x_{1t} such that the forecast of sales x_{2t+L} made at time t, $\hat{x}_{2t}(L)$, is zero(or target value). The optimal control strategy was written as:

$$\hat{x}_{2t}(L) = 0 \tag{2}$$

Since the observation $x_{2t+L} = \hat{x}_{2t}(L) + e_{2t}(L)$, the actual value = the forecasting + the forecasting error, It is clear that after implementing the control equation the output is given by $x_{2t+L} = e_{2t}(L)$ i.e., the actual value =the forecasting error.

i. Application Minimum Variance Control Strategy on Advertising-Sales Lydia Pinkham Company

The optimal control equation and effectiveness are presented in four steps:

A1. Model before control

Let us consider the model advertising-sales

$$\left.\begin{aligned} x_{1t} &= -0.132x_{1t-1} + 0.136x_{2t-1} + a_{1t} \\ x_{2t} &= 0.165x_{1-1t} + 0.321x_{1t-2}\ 0.062x_{1t-3} - 0.254x_{1t-4} \\ &\quad -0.10x_{2t-1} + 0.274x_{2t-2} + 0.058x_{2t-3} - 0.133x_{2t-4} \\ &\quad +0.22\ a_{2t-1} - 0.40a_{2t-2} + 0.7a2t\text{-}3 + a_{2t} \end{aligned}\right\} \tag{3}$$

Where: $\gamma_{11} = 21337$ and $\gamma_{22} = 7084.4$ are the error variance of x_{1t} and x_{2t} respectively and the output variance $\sigma^2 = 14595$

The Closed Loop system for Advertising x_{1t} and Sales x_{2t} is represented by the following figure.

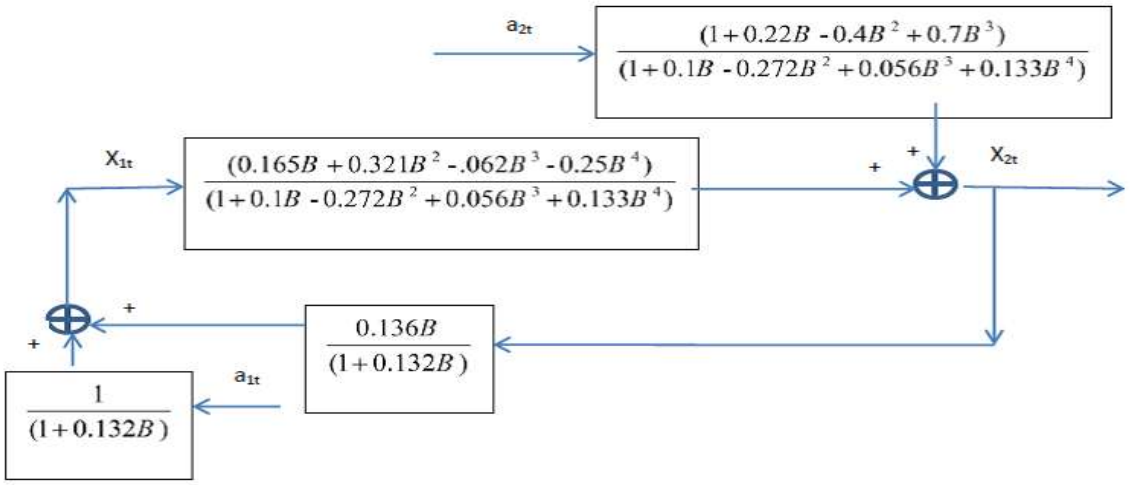


Figure (1): Advertising-Sales Block Diagram before Control

A2. Control equation from strategy

Because the lag is only after one sample, so to derive the control equation we need only the model for output x_{2t}, and we put the one ahead step conditional expectation equal to zero. Since L=1 then $\hat{x}_{2t}(1) = 0$

$$\hat{x}_{2t}(1)= 0.165x_{1t} + 0.321x_{1t-1} -0.062x_{1t-2} -0.254x_{1t-3}-0.10x_{2t}+0.274x_{2t-1} +0.058x_{2t-2} -0.133x_{2t-3} +0.22\, a_{2t} -0.40a_{2t-1} +0.7a_{2t-2} +0 = 0 \quad (4)$$

We can write the control equation as:

$$0.165x_{1t} + 0.321x_{1t-1} -0.062x_{1t-2} -0.254x_{1t-3}-0.10x_{2t}+0.274x_{2t-1} +0.058x_{2t-2} -0.133x_{2t-3} +0.22\, a_{2t} -0.40a_{2t-1} +0.7a_{2t-2} = 0 \quad (5)$$

The above control equation involves an a2t term that cannot be implemented directly; we need to represent a_{2t} term as an unknown parameter in x_{2t}.

A3. Model after control

We can write that:

$$0.165x_{1t-1} + 0.321x_{1t-2} -0.062x_{1t-3} -0.254x_{1t-4}-0.10x_{2t-1}+0.274x_{2t-2} +0.058x_{2t-3} -0.133x_{2t-4} +.22\, a_{2t-1} -0.40a_{2t-2} +0.7_{a2t-3} = 0 \quad (6)$$

Substitute from equation (6) in (4) for x_{2t} equation, we get optimal output equation as:

$$x_{2t} = a_{2t} \tag{7}$$

Now, if (6) substituted in (5), and divide all elements by 0.165 then the control equation for input x_{1t} is given by:

$$x_{1t} = \frac{(-0.72 + 0.763B - 4.59B^2 + 0.83xB^3)}{(1 + 2.0B - 0.378B^2 - 1.58B^3)} x_{2t} \tag{8}$$

The following figure shows the Block diagram after control

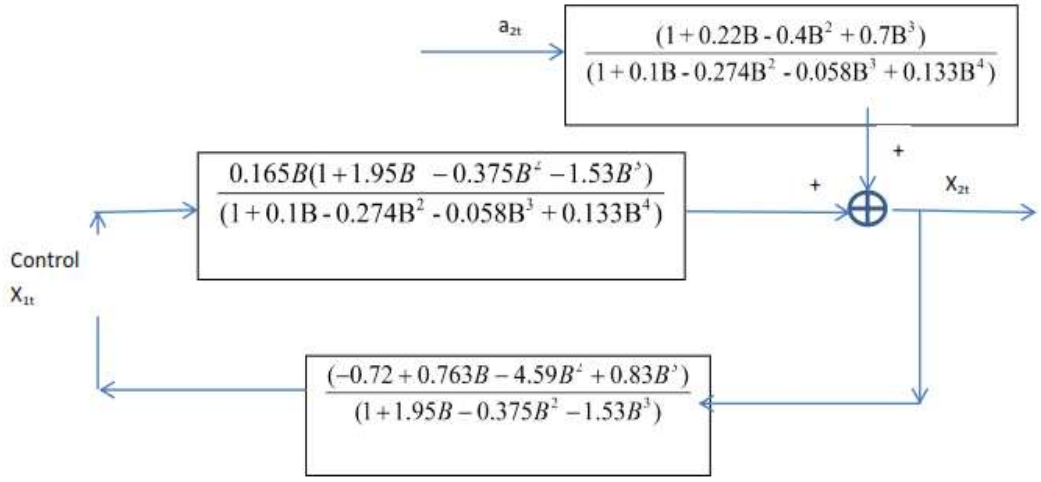


Figure (2): Minimum Variance Control Advertising-Sales Block Diagram

A4. Control efficient

Appling advertising-Sales minimum variance controls strategy (MVCS) appeared not efficient; we note that from minimum variance control advertising-sales block diagram in figure (2) the roots of denominator of feedback controller in the following transfer function:

$$\frac{(-0.72 + 0.763B - 4.59B^2 + 0.83B^3)}{(1 + 2.0B - 0.375B^2 - 1.53B^3)}$$

The polynomial $(1 + 2.0B - 0.375B^2 - 1.53B^3)$ has three roots (-1.68, -1.12, and 0.807). The first two roots -1.68, -1.12 lies outside of unit circle, and the same roots exist in the numerator of the transfer function model

$$\frac{0.165B(1 + 2.0B - 0.375B^2 - 1.53B^3)}{(1 + 0.1B - 0.272B^2 - 0.056B^3 + 0.133B^4)}$$

So, these roots canceled each other, this will cause the system unstable, these roots describe practical aspect in control called ***non-minimum phase***.

ii. Application Minimum Variance Incremental Control (MVIC) on Advertising-Sales Lydia Pinkham Company.

The process control schemes frequently specify an incremental control action whereby the control algorithm will output a signal sequence while it is actually applied to the system. The effect of this action is to introduce an integrator into the loop with the benefits that (i) transfer between controllers is possible, and (ii) Automatic steady-state reference set-point tracking occurs despite the presence of unmodelled disturbances Katsuhiko(2010),Wellstead and Zarrop(1991).

The idea of this control method is to modify the algorithm in which the control law generates with the control increment Δx_{1t} as in the figure (3) where:

x_{1t} : Advertising controlled variable

$\Delta x_{1t} = (x_{1t} - x_{1t-1})$: Increment of controlled variable

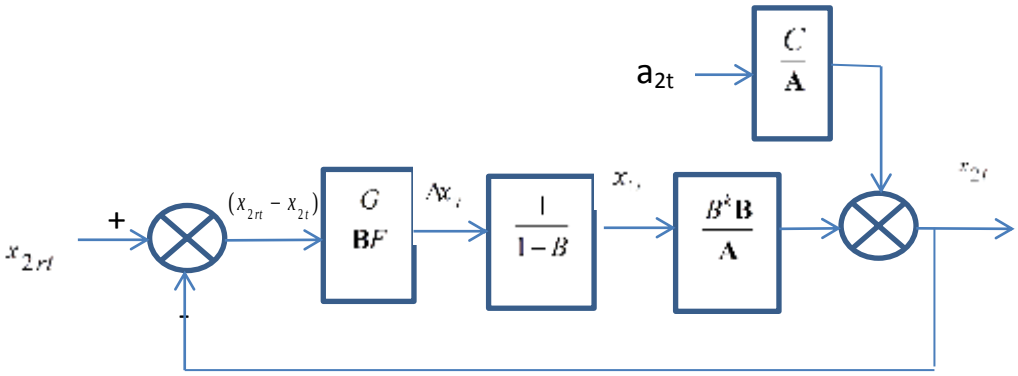


Figure (3): Minimum Variance Incremental Control (MVIC) Block Diagram

$\Delta x_{1t} = \frac{G}{\mathbf{B}F}(x_{2rt} - x_{2t})$, x_{2rt} is the reference sales

$x_{1t} = \frac{\Delta x_{1t}}{1-B}$, and $\frac{1}{1-B}$ as integrator

Where: $\mathbf{B}$, $\mathbf{A}$ and $\mathbf{C}$ are the polynomials of sales model. The closed loop equation becomes

$$x_{2t} = \frac{B^k \mathbf{B}}{\mathbf{A}} x_{1t} + \frac{C}{\mathbf{A}} a_{2t} \qquad (9)$$

B : back shift operator

k=1: the delay between x_{2t}, and x_{1t}

Where:

$$x_{2t} = \frac{B(0.165 + 0.321B - 0.062B^2 - 0.254B^3)}{(1 + 0.10B - 0.274B^2 - .058B^3 + 0.133B^4)} x_{1t} + \frac{(1 + 0.22B - 0.40B^2 + .7B^3)}{(1 + 0.10B - 0.274B^2 - 0.058B^3 + 0.133B^4)} a_{2t} \quad (10)$$

Controller polynomials are:

$$F = 1 + f_1 B + \ldots + f_{n_f} B^{n_f}$$
$$G = g_0 + g_1 B + \ldots + g_n B^{n_g} \quad (11)$$

Where:

$n_f = k - 1,\ k = 1,\ n_g = \max(n_a, n_c - k)$, n_a the order of polynomial **A**

The controller polynomials are obtained by solving the modified identity:

$$C = FA(1 - B) - B^k G \quad (12)$$

After applying the control equation we obtain the results:

Table (1): Comparing the Advertising Sales System With and Without Control

Mode	Advertising	Sales	Desire
Without Control	838	1052.000	---
	994	1102.000	---
	1020	1355.000	---
	865	1323.000	---
	819	1296.000	---
	83	1127.000	---
	56	1170.000	---
	224	1059.000	---
	881	1116.000	---
	436	1214.000	---
	160	966.0000	---
	68	1089.000	---
	Sum = 6444	**Sum=13869**	---
With Control	747.9677	1031.793	1149.400
	1254.337	1101.992	1253.000
	1146.545	1247.791	1563.200
	1583.104	1336.489	1505.400
	751.7563	1313.515	1373.200
	1446.312	1294.322	1258.600
	370.1893	1361.628	1224.000
	2632.395	1549.800	1245.200
	464.5491	1722.450	1350.800
	4748.497	1756.547	1498.000
	-2323.859	1643.835	1260.600
	8056.888	1448.757	1182.400
	Sum= 20878.681	**Sum=16808.92**	**Sum=15863.80**

The results show that the advertising appeared large and negative for some values. This indicates that failing of (IC) control for our reference that has large changes. From the results review, we conclude that a one control aspect indicates that the incremental minimum variance control is suitable if the reference is constant or changes slowly then the control can be tracked the system.

iii. Generalized Minimum Variance Controller (GMVC)

The MVC approach can be extended to GMVC control which controlled in the system output and to follow the desired reference $r(t)$ varying with time, and to solve the problem of non-minimum phase or zero cancellation Katsuhiko(2010), Wellstead and Zarrop(1991). For the above two applications, we applied Generalized minimum variance controller GMVC method to solve the problem of changing the targets of adverting- sales system. We will see that after applied GMVC that the variance efficiency will improved by (63%). The design of GMVC controller beside solve the problems of MVC which improve the characteristic of the model, it is contribution in solve the problem of ***non-minimum phase***, and general reference (constant or vary with time).

The system model transfer function can be written as:

$$x_{2t} = \frac{\mathbf{B}}{\mathbf{A}} x_{1t} + \frac{C}{\mathbf{A}} a_{2t}$$

Where *B* is back shift operator with power L as delay time, with system polynomials are B, A and C.

Because x_{1t} the will only affect x_{2t} for $t \geq t + L$, i.e. the first output samples affected by Instant input after L samples, the system model that appears in the cost function is:

$$x_{2t+k} = \frac{\mathbf{B}}{\mathbf{A}} x_{1t} + \frac{C}{\mathbf{A}} a_{2t+k} \tag{13}$$

The approach introduces a system pseudo-output $\phi(t)$ described by:

$$\varphi_{t+k} = P\mathrm{x}_{2t} + \mathrm{Q}\,\mathrm{x}_{1t} - \mathrm{R}\,\mathrm{r}_{t} \tag{14}$$

The main of GMVC is to minimize the $\phi(t)$ function or the cast function. The system as in figure (4) the output based on the original system with feed forward added to x_{1t} with filter Q, to x_{2t} with filter P, and to set point $r(t)$ with filter R. Feedforward terms (Q; P) are to shift the system open loop zeros from to (P+ Q) where $\mathbf{B}$, $\mathbf{A}$ are the polynomials of the system Guenfaf and Azira(2016), Yanou et al.(2014), Jelali(2012), Katsuhiko(2010), Mohamed et al.(2009).

Substitute equation (13) in equation (14) we get:

$$\begin{aligned}\phi_{t+k} &= Px_{2t+k} + Qx_{1t} - Rr_t \\ &= P\left(\frac{\mathbf{B}}{\mathbf{A}}x_{1t} + \frac{C}{\mathbf{A}}a_{2t} + Qx_{1t} - Rr_t\right) \\ &= \frac{P\mathbf{B}}{\mathbf{A}}x_{1t} + \frac{PC}{\mathbf{A}}a_{2t+k} + Qx_{1t} - Rr_t \\ &= \left(\frac{P\mathbf{B}}{\mathbf{A}} + Q\right)x_{1t} + \frac{PC}{\mathbf{A}}a_{2t+k} - Rr_t \\ &= \left(\frac{P\mathbf{B} + Q\mathbf{A}}{\mathbf{A}}\right)x_{1t} + \frac{PC}{\mathbf{A}}a_{2t+k} - Rr_t\end{aligned} \tag{15}$$

The cost function to be minimized is the variance of pseudo-output:

$$J = E[\phi^2_{t+k}] \tag{16}$$

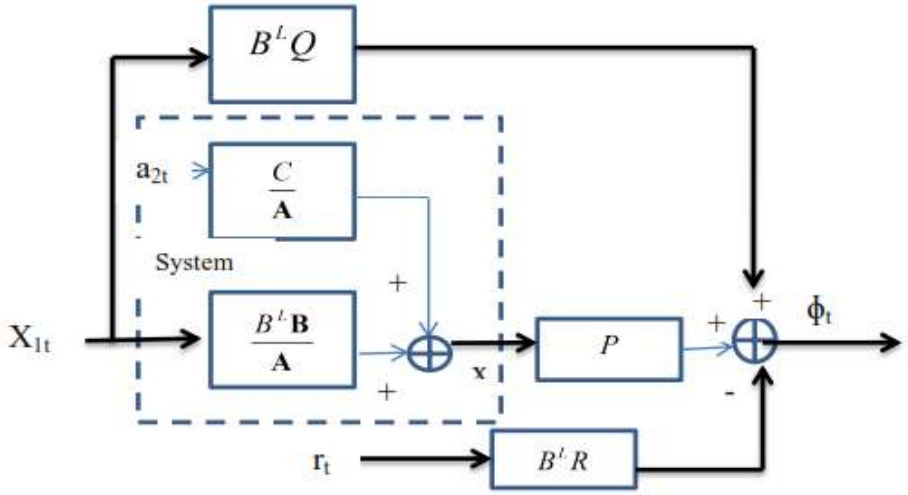


Figure (4): Generalized Minimum Variance Control (GMVC) Block Diagram

iv. Design GMVC Controller Strategy

Begin by defining E and G via identity as the following:

$$PC = E\mathbf{A} + B^k G \tag{17}$$

Where:

I- The simple design put p=1

II- $E = 1 + e_1 B \dots \dots \dots e_{n_e} B^{n_e}$, $n_e == k - 1$

III- $G = g_0 + g_1 B \dots \dots \dots g_n B^{n_g}$

IV- $n_g = max(n_{a-1}, n_p + n_c - k)$, n_a is the order of polynomial **A**

Multiply the system equation (13) by E**A** to yield:

$$E\mathbf{A}x_{2t+k} = E\mathbf{B}x_{1t} + ECa_{2t+k} \quad (18)$$

from equation (18) substitute by EA in equation (17) we get:

$$PCx_{2t+k} = E\mathbf{B}x_{1t} + Gx_{2t} + CEa_{2t(t+k)} \quad (19)$$

Add the term $(QCx_{1t} - CRr_t)$ to both sides of equation (19) we get:

$$C(\varphi_{t+k}) = (E\mathbf{B} + QC)x_{1t} + Gx_{2t} - CRr_t + CEa_{2t+k} \quad (20)$$

This can be cast in the form:

$$\phi_{t+k} = \frac{1}{C}[(E\mathbf{B} + QC)x_{1t} + Gx_{2t} - CRr_t] + Ea_{2t+k} \quad (21)$$

The error term $Ea_{2t(t+k)}$ is uncorrelated with the remainder of the right-hand side. The cost function is therefore minimized by setting the first term on the right-hand side equal to zero, the GMVC controller given by:

$$(E\mathbf{B} + QC)x_{1t} = -Gx_{2t} + CRr_t \quad (22)$$

or the control equation become in the form:

$$Fx_{1t} + Gx_{2t} - Hr_t = 0 \quad (23)$$

Where:

$$F = E\mathbf{B} + QC$$
$$H = CR, R = p(1) = 1$$

From the control equation (23) we get:

$$x_{1t} = \frac{-G}{F}x_{2t} + \frac{H}{F}r_t \quad (24)$$

Substitute with equation (24) in system equation (14) we obtain:

$$x_{2t+k} = \frac{\mathbf{B}}{\mathbf{A}}x_{1t} + \frac{C}{\mathbf{A}}a_{2t+k} \quad (25)$$

Then the closed loop equation become in the form:

$$x_{2t} = \frac{B^k\mathbf{B}H}{\mathbf{A}F + B^k\mathbf{B}G}r_t + \frac{FC}{\mathbf{A}F + B^k\mathbf{B}G}a_{2t} \quad (26)$$

Using the following relations:

I- $F = E\mathbf{B} + QC$

II- $H = CR$

III- $PC = E\mathbf{A} + B^{k}G$

in the closed loop equation (26).

$$x_{2t} = \frac{B^{k}\mathbf{B}R}{(P\mathbf{B} + Q\mathbf{A})} r_t + \frac{(\mathbf{B}E + QC)}{(P\mathbf{B} + Q\mathbf{A})} a_{2t} \qquad (27)$$

This is an alternative form of the closed loop equation (26).

Note:

I- If Q=0 the system return to MV with zero cancellation.

II- The algorithm is equivalent to a pole assignment algorithm if the poles defined by roots of a chosen polynomial T are used to fix, P, and Q through the identity PB+ QA = T. For simple design chose P, and Q =1, and B+A = T if stable roots.

v. The Closed Loop of the GMVC Block Diagram

The control equation for the GMVC controller is as in equation (23).So final control equation: $x_{1t} = \frac{-G}{F} x_{2t} + \frac{H}{F} r_t$ represented by the block diagram as in the following figure:

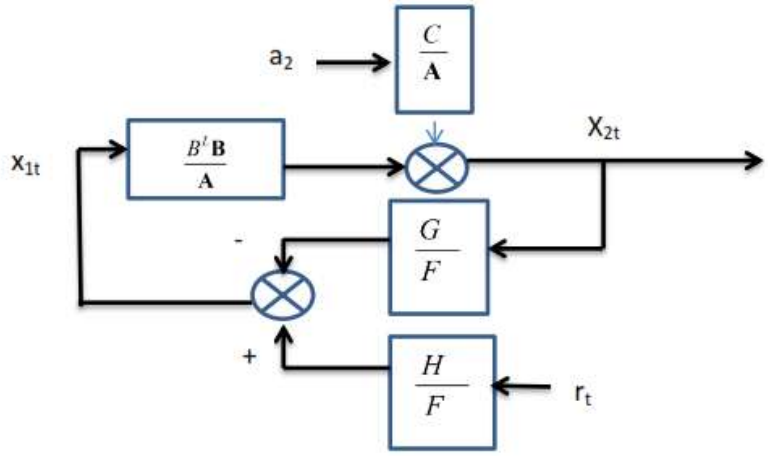


Figure (5): Closed Loop GMVC Block Diagram

If we simplified this block diagram we reached to the diagram which represents the closed loop models:

$$x_{2t} = \frac{B^{k}\mathbf{B}\mathrm{H}}{\mathbf{A}F + B^{k}\mathbf{B}G}\mathrm{r}_{t} + \frac{FC}{\mathbf{A}F + B^{k}\mathbf{B}G}a_{2t} \quad (28)$$

Or:

$$x_{2t} = \frac{B^{k}\mathbf{B}R}{P\mathbf{B} + Q\mathbf{A}}\mathrm{r}_{t} + \frac{(\mathbf{B}E + QC)}{P\mathbf{B} + Q\mathbf{A})}a_{2t} \quad (29)$$

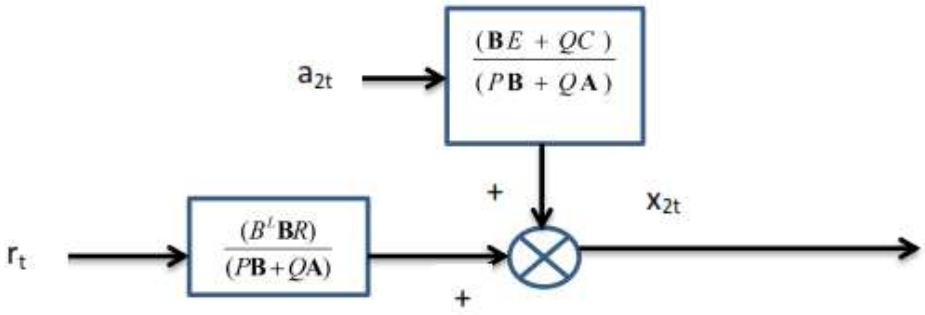


Figure (6): Simplified Closed Loop GMVC Block Diagram

For advertising-sales Lydia Pinkham system, using the desired sales time series to design and control the sales during the next one year. The control equation describes the relationship between advertising and sales time series to control sales through advertising expenditure. Figures (7) describe the control process corresponding to the application case.

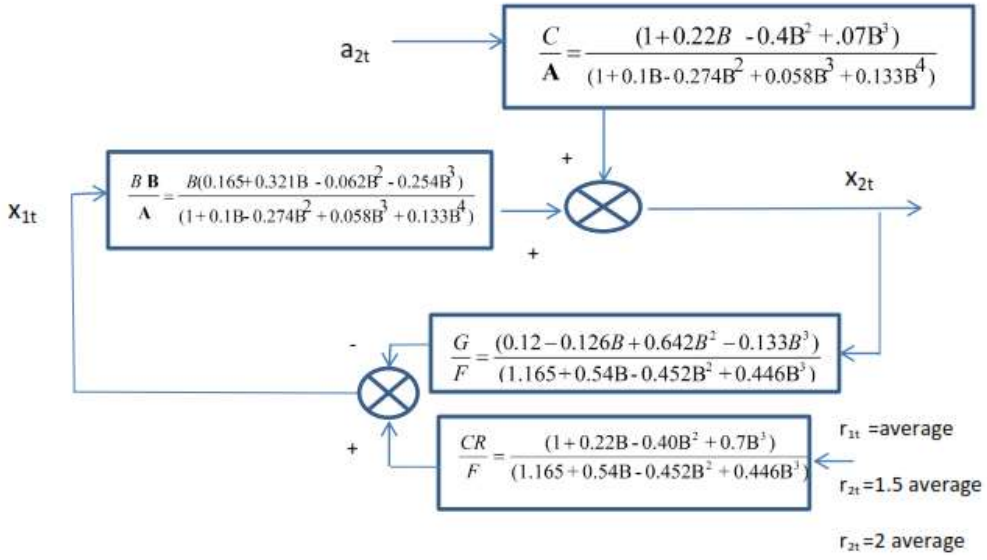


Figure (7): GMVC Closed Loop Block Diagram advertising-Sales after Control

We can simplified GMVC Closed Loop Block Diagram advertising-Sales after Control in figure (7) as:

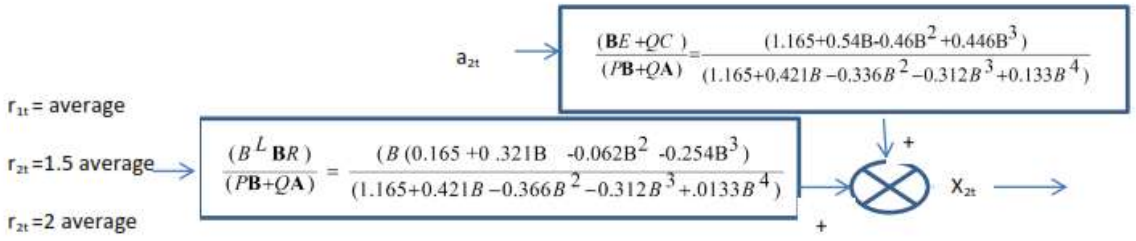


Figure (8): Simplified GMVC Closed Loop Block Diagram advertising-Sales after Control

I- Application Generalized Minimum Variance Control (GMVC) on Advertising-Sales Lydia Pinkham Company: The Desired Sales Taken as the Average Values for Each Month for Five Years

Table (2): Comparing the Advertising Sales System With and Without Control

Mode	Advertising	Sales	Desire
Without Control	838	1052.000	---
	994	1102.000	---
	1020	1355.000	---
	865	1323.000	---
	819	1296.000	---
	83	1127.000	---
	56	1170.000	---
	224	1059.000	---
	881	1116.000	---
	436	1214.000	---
	160	966.0000	---
	68	1089.000	---
	Sum = 6444	**Sum=13869**	---
With Control	857.9196	1136.737	1149.400
	944.3612	1348.572	1253.000
	1292.847	1699.696	1563.200
	908.8084	1621.222	1505.400
	555.6404	1330.359	1373.200
	106.0181	1168.735	1258.600
	295.7551	1235.814	1224.000
	276.4477	1275.266	1245.200
	928.1672	1301.189	1350.800
	688.3220	1437.353	1498.000
	291.8915	1325.448	1260.600
	36.92023	1192.349	1182.400
	Sum=7183.099	**Sum=16072.74**	**Sum=15863.80**

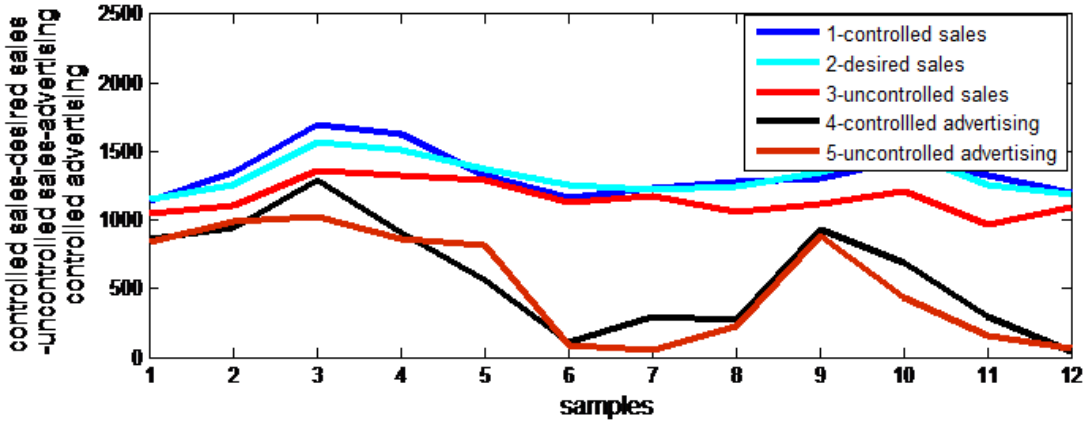


Figure (9): Advertising and Sales System Under Control

The figure shows five types of graph lines, first line is the controlled sales, the second line is the desired sale, the third line represented the uncontrolled sales, the fourth line represented the controlled advertising, and the fifth line is uncontrolled advertising. We study the variance efficiency after applied GMVC. The design of the GMVC controller contributed to solve the problems of MVC to intended the improve the characteristic of the model, and appear professional in solving a non-minimum phase or called zero cancellation control practical problem by manipulating with the general reference, special in the pattern of reference which changes across time.

Table (3): Comparing Minimum Variance Control Methods for Advertising-Sales System With and Without Control

Method	Without Control	With Control
IC	Net uncontrolled Sales=7425 The uncontrolled model variance=14595	▪ Advertising results shows negative and large values. ▪ This indicates that failing of this control for our reference that has large changes.
GMVC- First Case		▪ Net Control Sales=8889 ▪ Increasing in sales=11% ▪ Variance of desired sales=5372 ↓ ▪ Variance improvement=63%
GMVC- Second Case		▪ Net Control Sales=12666.62 ▪ Increasing in sales=38% ▪ Variance of desired sales=13652 "↑ ▪ Variance improvement=6.4%
GMVC- Third Case		▪ Net Control Sales=16043.6 ▪ Increasing in sales=61% ▪ Variance of desired sales=30924 ↑↑ ▪ Variance non-improvement=110%

Conclusion

The optimal control is defined as those adjustments input values that yield minimum variance or minimum mean squared error of the output, where the reference is either zero or constant. Table (3) shows the results for different minimum variance control methods; the advertising-sales system is varying at each sample, so we use incremental control (IC) which based on introduced integrator in the controller law. The rustle in the table (1) indicated that failing in incremental control because advertising results with control shows large and some negative values which indicate that failure of this control for our reference that has large changes and concludes that the (IC) is more appropriate for the reference system with change slowly. The analysis of the results for the three cases of different references for desired sales given in table (3) shows that the GMVC control has robust results and made the output follow the desired. In the first case, the advertising-sales Lydia Pinkham company under controlled planning achieves profit than that uncontrolled with consideration the criteria of the accurate measurement for the controller is the minimum variance of the output deviation from the desired references. The first desired reference taken as the average of each month during five years gives a (63%) improving percentage in the variance. The second case takes the reference value one and half of this value gives a (6.4%) improving percentage in the variance. The third case, although of achieving some profit but the deviation of the variance is less than that uncontrolled, so we chose the references that attainment the minimum variance of the system which described by the reference in the first case application as the recommended in the literature Alipouri and Poshtan (2016), Donoghue S (2008).

References:

Alipouri, Y. and Poshtan, J. (2016). Design of robust minimum variance controller using type-2 fuzzy set. Transactions of the Institute of Measurement and Control, 38(3):315–326.

Athans, M. and Falb, P. L. (2013). Optimal control: an introduction to the theory and its applications. Courier Corporation.

Bhattacharyya, M. N. (1982). Lydia Pinkham data remodelled. Journal of time series analysis, 3(2):81–102.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

Donoghue S, J.W. Finch, D. G. J. A. (2008). Generalized minimum variance controller as a velocity loop controller of a casting drum drive in a polyester manufacturing line. Proceedings of the World Congress on Engineering, WCE 2008, July 2 - 4, 2008, London, U.K., Vol III

Feichtinger, G., Hartl, R. F., and Sethi, S. P. (1994). Dynamic optimal control models in advertising: recent developments. Management Science, 40(2):195–226

Guenfaf, L. and Azira, M. (2016). Generalized minimum variance control for mdof structures under earthquake excitation. Journal of Control Science and Engineering, 2016.

Hanssens, D. M., Parsons, L. J., and Schultz, R. L. (2003). Market response models: Econometric and time series analysis, volume 12. Springer Science & Business Media.

Holly, S. and Zarrop, M. (1979). Optimal control for econometric models: An approach to economic policy formulation. Springer.

Jelali, M. (2012). Control performance management in industrial automation: assessment, diagnosis and improvement of control loop performance. Springer Science & Business Media

Katsuhiko, O. (2010). Modern control engineering. Prentice Hall, New Jersey. (5th Edition).

Kirk, D. E. (2012). Optimal control theory: an introduction. Courier Corporation.

Lewis, F. L., Vrabie, D., and Syrmos, V. L. (2012). Optimal control. John Wiley & Sons.

Mohamed, S., Zayed, A., and Abolaeha, O. (2009). New feed-forward/feedback generalized minimum variance self-tuning pole-placement controller. In Proceeding of World Academy of Science, Engineering and Technology, pages 998–1001.

Pandit, S. M., Wu, S.-M., et al. (1983).Time series and system analysis with applications. Wiley New York.

Rufino, C. (2008). Lagged Effect of TV Advertising on Sales of an Intermittently Advertised Product. *DLSU Business & Economics Review*, *18*(1), 1-12.

Seierstad, A. and Sydsaeter, K. (1986). Optimal control theory with economic applications. Elsevier North-Holland, Inc.

Sethi, S. P. and Thompson, G. L. (2006). Optimal control theory: applications to management science and economics. Springer Science & Business Media.

Todorov, E. and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. Nature neuroscience, 5(11):1226.

Wei, W. W. et al. (2006). Time series analysis: univariate and multivariate methods. Pearson Addison Wesley.

Wellstead, P. and Zarrop, M. B. (1991). Self-tuning systems: control and signal processing. John Wiley & Sons, Inc.

William W. S. Wei. (2019). Multivariate Time Series Analysis and Applications. Wiley Series in Probability and Statistics

Yanou, A., Minami, M., and Matsuno, T. (2014). Design method of generalized minimum variance control using strong stability rate.
URL:https://www.semanticscholar.org/paper/Design-Method-of-Generalized-Minimum-Variance-Using-Yanou-Minami/2bd79ea88a30cbfd96e94fff5b63c6505e28941f

Chaotic Analysis of Egyptian Stock Market with the Application to EGX 30 Price Index

Eman Talaat Morad [1], El-Houssainy A. Rady [2] and Ayman A. Amin [3]

[1] Teaching Assistant, Department of Business Administration, Faculty of Commerce, Cairo University, Egypt

[2] Professor, Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

[3] Assistant Professor, Department of Statistics, Faculty of Commerce, Menoufia University, Egypt

Abstract

Chaos theory as one of the nonlinear dynamic tools can be useful in explaining the dynamics of financial markets, since chaotic models are capable of exhibiting behaviors similar to those observed in financial data. With the aim of this paper to apply the chaos theory to the Egyptian stock market, the researcher first presents the basic features of chaos theory, as well as, the rationales for bringing chaos theory to the attention of financial researchers. Accordingly, the paper studies the stochastic characteristics of the EGX30 Egyptian stock market index, and evaluates whether its movements can be explained through chaotic models. In particular, the EGX30 index is analyzed in the time and frequency domains, where the series is found to exhibit serial autocorrelation and deviations from normality with fat tails in their distributions. In addition, the Brock-Dechert-Scheinkman (BDS) test is used, a powerful test for independence, to demonstrate that the EGX30 index has nonlinear components. This result was compatible with a chaotic explanation, so the researcher proceeds further with the R/S test, which reveals a long-term memory and fractal structure of the EGX30 index, even in the presence of noise.

Keywords: Chaos, BDS Test, Nonlinearity, Rescaled Range (R/S) Analysis, Hurst Exponent.

1. Introduction

The mechanisms that explain the dynamics of asset prices and their returns are an essential issue in finance. The traditional approach supposes that fluctuations in these variables are hugely the result of random processes that can be represented by linear stochastic models. Indeed, this approach has been experimentally contested and there has been an increasing interest in nonlinear approaches and different types of nonlinear stochastic specifications (Karytinos (1999); Chandra, and Thenmozhi (2017)).

According to Harriff (1997), in recent years, the theories of nonlinear systems generally, and chaotic systems particularly, have received a major deal of interest in the financial community. A "chaotic" process in this context is one that generates time paths of variables of interest that give the appearance of being generated by some random process. Historically, the study of complex dynamics has its roots in the physical

sciences. However, the idea of the presence of at least some deterministic component to the price dynamic has great attractiveness to those who are willing to predict the time paths of market prices of investment assets. Such determinism may be susceptible to being uncovered through more sensitive and sophisticated methodologies, although it may be out of reach to most conventional statistical tests constructed on linear hypotheses. To the extent that nonlinear dependence exists within or across asset price time series, an understanding of the form of the underlying mechanism can lead to a greater degree of predictability, compared to the case of conventional statistical models.

As explained by Karytinos (1999), the question of whether chaos theory may be proven to be an important mathematical tool in getting a better understanding of the financial markets is still unresolved. Many authors have employed methods and techniques from the nonlinear dynamics field to analyze different economic and financial series and attempt to determine whether chaotic behavior is present. Up to now, most of empirical findings have been contradictory, and the source of this controversy includes: (i) Nature of employed data sets (namely sample size, noise level, aggregation method, and so on), (ii) Different methods and techniques that have been used, (iii) Improper use of the chaotic testing framework, and (iv) The fact that most chaotic techniques are not statistical, which makes the interpretation of the results more difficult (Faggini and Parziale (2012); Aguirre and Letellier (2009)).

In this paper, the researcher studies the stochastic characteristics of the EGX30 Egyptian stock market index, and evaluates whether its movements can be explained through chaotic models. In particular, first the EGX30 series in the time and frequency domains is analyzed, where the series is found to exhibit serial autocorrelation and deviations from normality with fat tails in their distributions. Second, the BDS test is used, a powerful test for independence, to demonstrate that the series has nonlinear components. This result was compatible with a chaotic explanation, so the researcher proceeds further with the R/S test, which reveals a long-term memory and fractal structure of the series, even in the presence of noise.

The rest of this paper is structured as follows. Section 2 discusses the background of chaos theory, and section 3 presents the data and methodology, and then discusses the results of BDS test and R\S analysis. While a summary of the findings and the conclusions of this research are presented in Section 4.

2. The Concept of Chaos

Many researchers such as Karytinos (1999), Baumol and Quandt (1985) indicated that in the last few years, the exploding of informatics developments and computing ability have made possible the in-depth exploration of complex systems especially in the Natural Sciences. One of these complex systems that have been discovered is nonlinear systems, and the exploration of nonlinear systems that aim at a new understanding of the laws of the disorder is called "chaos theory" or "nonlinear dynamics". More precisely, Chaos is a subset of the more general nonlinear dynamics framework.

As explained by Gleick (1987) chaos is a "new science" which, for some of its advocates, is equally revolutionary to relativity and quantum mechanics. It can be considered as the latest scientific revolution or using as a "paradigm shift" from the still prevailing in most scientific fields "linear paradigm" to the "nonlinear paradigm".

According to some authors, for example, Gunaratne (2003) and Barton (1994), Chaos can be considered as a generalized paradigm in nature.

Karytinos (1999) mentioned in his study that Chaos theory has been developed in the theoretical framework of dynamical systems analysis and drew heavily on specialized disciplines like differential topology, fractal geometry, and so on. It cuts across the interdisciplinary dichotomy between "determinism" and "randomness", which characterizes the way that traditional science approaches the various phenomena. Actually, these two opposite modes of behavior seem to reconcile in the chaos theory framework.

The development of Chaos theory is an interdisciplinary effort. Therefore, it is not surprising that it has been proven to be very useful in analyzing the behavior of many phenomena, the study of which has been the subject of a wide range of different disciplines, such as physics, biology, meteorology, chemistry, medicine, ecology and economics and finance (Oestreicher (2007); Selvam (2012)).

According to Karytinos (1999), the modern study of chaos dates back to the 1960s. In a brief reference, the most important theoretical developments of Chaos theory can be listed as follows:

- In 1963, Lorenz, a meteorologist who rediscovered the Sensitive Dependence on Initial Conditions (SDIC) principal in a simple non-linear dynamical model for weather prediction consisting of three differential equations that produce the famous "Lorenz Attractor".
- In 1963, Smale, a mathematician specialized in topology who was the first to connect topology with Chaos theory. His innovation was a topological transformation known as "horseshoe", which provides a base for understanding the chaotic properties of dynamical systems.
- In 1975, May, a mathematician and a biologist respectively who studied a very simple but now the most famous chaotic system known as the "Logistic Equation" or the "Logistic Map". This simple system has been extensively used in the literature for educational and modeling purposes. It should be also mentioned that York was the one who gave to Chaos its very name.
- In 1978, Feigenbaum discovered universal laws hidden in the routes to Chaos. Feigenbaum proved that there are structures in non-linear systems that are always the same, which means that chaotic routes obey the same universal laws.
- To close this short chain of the early discoveries related to Chaos, it is necessary to mention Mandelbrot in 1982, who developed the notion of fractals or self-similarity across scales. This concept plays an important role in the description of "Attractors", the basic dynamical object under consideration in the chaotic framework.

3.1 Data

The data employed in this paper is the EGX 30 daily Egyptian stock market index downloaded from the Egyptian Stock Exchange website (egx.com.eg). The available data for this index allows us to use more than 4393 daily observations during the sample period between January 1998 and December 2015. The daily returns are calculated using the log difference of the daily closing prices of EGX 30 index selected, as follows:

$$R_t = \ln\left(\frac{p_t}{p_{t-1}}\right) = \ln(p_t) - \ln(p_{t-1})$$

Where R_t and p_t are the daily returns and daily closing price of the stock index at time t, respectively. In this analysis, the Eviews 8.1 and Matlab R2016a are used for Windows.

3.2 Analysis and Results

3.2.1 Statistical Description

First, the basic statistical properties of EGX 30 daily returns are analysed using descriptive statistics as presented in Table (1).

Table 1. Descriptive analysis of the daily returns of the EGX 30 index

Indices	Statistical measure				
	Sample size	**Median**	**Mean**	**SD**	**Probability**
EGX 30	4393	.000645	.000443	0.017319	
	Skewness	**Kurtosis**	**Interquartile range**	**Jarque-Bera**	0.000000
	-.350851	11.75502	0.01764	14120.34	

Skewness: The coefficient of asymmetry is negative for the EGX 30 index, which indicates that the distribution of daily returns for the index is asymmetric to the left.

Kurtosis: Daily returns for EGX 30 index have a high kurtosis much greater than 3, and this reflects the presence of a leptokurtic distribution.

The leptokurtic distribution has heavier tails or a higher probability of extreme outlier values when compared to a normal distribution. A leptokurtic distribution means that the investor can experience broader fluctuations resulting in greater potential for extremely low or high returns, so valuation models for prices can generate errors if it is assumed that the prices follow a normal distribution.

In **Jarque-Bera test**, the null hypothesis of normality is rejected. The test confirms that the returns of the EGX 30 index are not normally distributed.

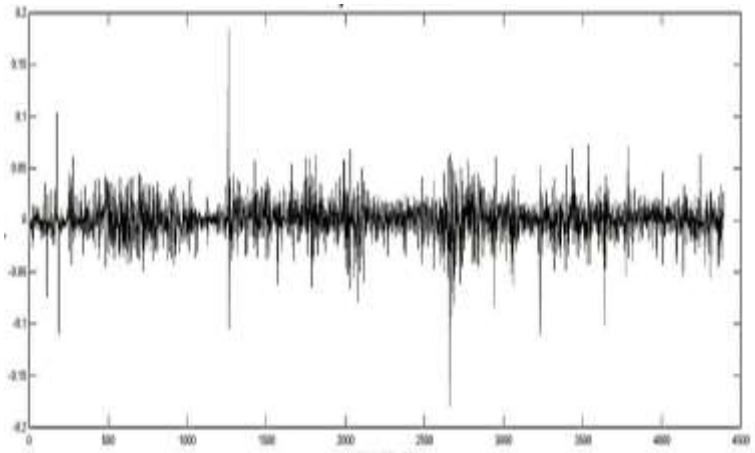


Fig. (1): The returns-time series plot of EGX30

Table (1) and Figure (1) together shows that the EGX 30 index returns are leptokurtic and skewed and accordingly deviate from the normal distribution, this matter is very common in the stock market. Also, Figure (1) shows the variation clustering in the daily stock returns.

Finally, other important conclusions from this figure can be drawn, they are as follows: a simple visual examination of the series indicates that yields display specific properties of non-linear models: volatility clustering, heteroscedasticity phenomenon, and significant leverage effects.

It is noted that the volatility is focused in short periods and this indicates possible correlations between current volatilities and historical volatilities, successive periods of high-volatility followed by large variations with periods of low-volatility followed by small variations in yields.

3.2.2 Stationarity Tests

To check and analyze the hypothesis of stationarity of the daily return series, unit root tests are implemented to identify and determine the order of integration using the stationarity tests: Augmented Dickey-Fuller and Phillips-Perron. The results of these tests are listed in Table 2.

Table 2. ADF and PP test results

Augmented Dickey Fuller Test (ADF)						
Indices	H	P	t-statistic	Critical value (1%)	Critical value (5%)	Critical value (10%)
EGX30	1	0.001	-54.62546	-3.431655	-2.862002	-2.567059
Phillips Perron Test (PP)						
Indices	H	p	t-statistic	Critical value (1%)	Critical value (5%)	Critical value (10%)
EGX30	1	0.001	-54.81730	-3.431655	-2.862002	-2.567059

Stationarity tests have been used in the null hypothesis that the daily return series analyzed contains a unit root, which means that it is not stationary.

As it seems from table 2, for all the daily return series, the statistic test has a value lower than the critical value at all levels of significance, which indicates that the hypothesis of a unit root is rejected; therefore, daily return series selected in the analysis are stationary, it can else be said that they are integrated of order 0.

In financial series, an important problem appears; that is the serial correlation of residuals. Wherefore in the current study, the researcher checks if there is a correlation in each return series selected.

At first, autocorrelation functions and partial autocorrelation functions are analyzed estimated up to lag 15 as presented in figures (2) & (3).

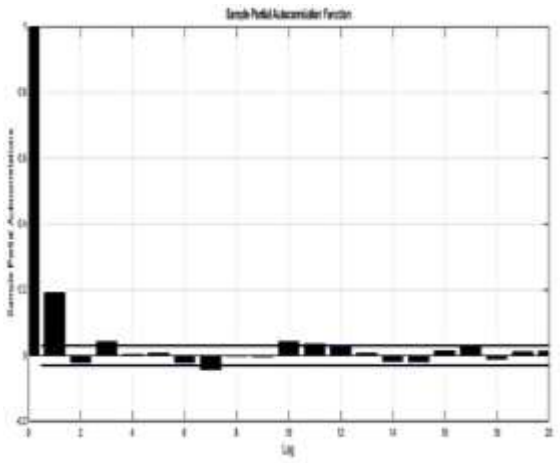


Fig. (2): EGX30 partial autocorrelation

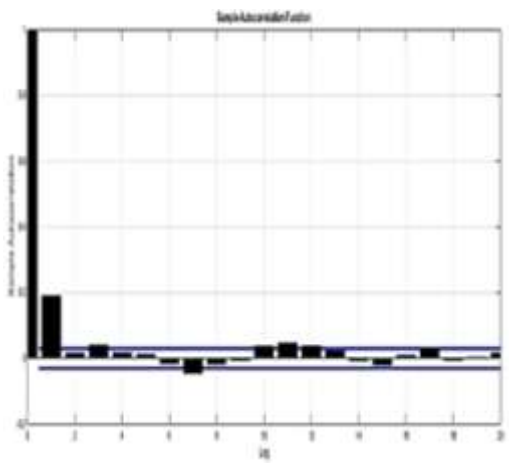


Fig. (3): EGX30 autocorrelation

Figure (2) and Figure (3) for the series indicate that daily returns in the market are linearly dependent.

It is clear that the boundary lines in the above plot are compatible with 1% significance level. As illustrated, autocorrelation is very strong at least for the first lag, but there is a more marginally significant autocorrelation coefficient in longer lags for the series.

Autocorrelation, which is pronounced in the series, is statistically significant for the first 15 lags, and partial autocorrelation indicates an AR (2) for the EGX30.

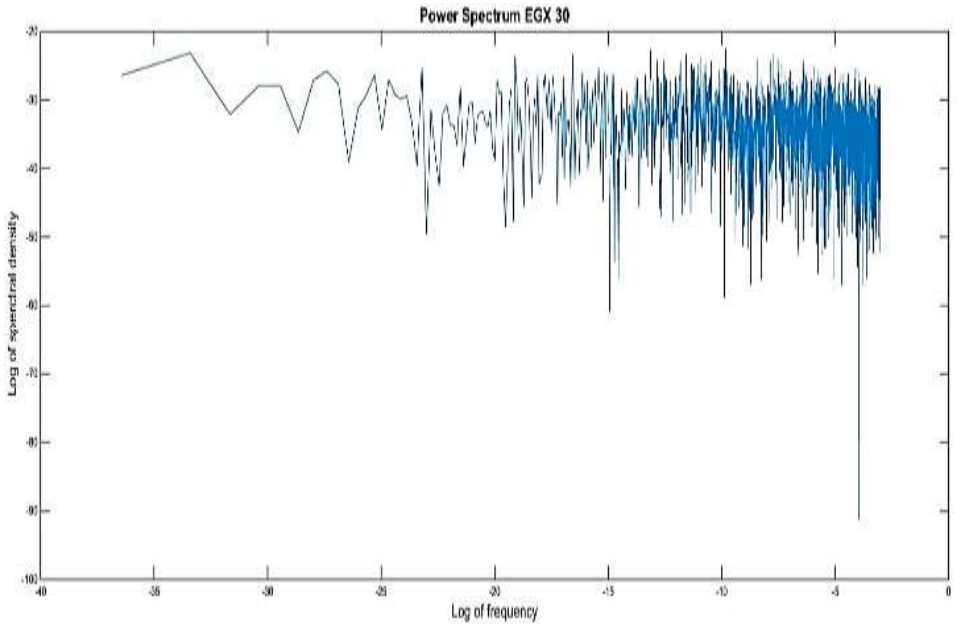


Fig. (4): Power Spectrum of EGX30 index

In figure (4), the power spectrum of the EGX30 daily return reveals two different regions. The left region that is horizontal to the frequency axis, indicates the strong presence of noise. Whilst the right region presents a decline, indicative of power-law behavior and a nonwhite noise structure.

3.2.3 Variance Ratio Tests of Random Walk Hypothesis

In the financial time series, some evidence appears that refers to deviations from normality and time-varying volatility, so it is important to perform a test that is robust to heteroscedasticity.

The variance ratio (VR) test of the random walk is conducted. The main property of the random walk hypothesis that the variance of the residual variable must be a linear function of the time.

Given the RW1:$r_t = \mu + \varepsilon_t$, as returns are (IID), so that var[r_t+r_{t-1}]=2var[r_t]. Thus, it is possible to define whether the random walk hypothesis (RWH) is rational by checking variance ratio: $\mathrm{VR}(2) = \frac{\mathrm{var}[r_t + r_{t-1}]}{2\mathrm{var}[r_t]}$

Then, to accept the random walk hypothesis, the variance ratio must be equal to one.

As can be seen from Table 3, the variance ratio for the return series is less than 1 at the confidence level of 1%. Hence, based on the VR test, the null hypothesis is rejected that the series would follow a random walk model. This means that the daily return series can follow some predictable patterns.

Table 3. VR test results

	Value	VR at period 2	VR at period 4	VR at period 8	VR at period 16
EGX 30	12.97722	0.608228	0.303588	0.157507	0.076868
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

This rejection of the IID-null hypothesis of daily return observations refers to either an underlying a nonlinear stochastic process or chaotic process. In the next step, the results from the BDS test are discussed to check the nature of dependence current in the data. In our analysis, the BDS test is used to detect nonlinear structure in the data, since the nonlinear structure is consistent with a chaotic explanation.

3.2.4 BDS Test for Nonlinearity

3.2.4.1. Description of BDS Test

A powerful test used for analysis independence and nonlinear dependencies for data was developed by Brock, Dechert, and Scheinkman in 1996 and was based on the correlation integral. The BDS test can only be used to produce indirect evidence about nonlinear structure in data because the sampling distribution of the test statistic is unknown. The BDS test is a useful statistic to test for patterns that occur less (or more) often than would be expected in independent data. This statistic can be used to test for the remaining nonlinear dependence in the ARIMA process residuals. The null hypothesis of this test may be formulated as (see Afonso and Teixeira (1998); Bhattacharya and Sensarma (2013)):

$\boldsymbol{H_0}$: Sample observations generated by independently and identically distributed (I.I.D.)

$\boldsymbol{H_1}$: Non-linear dependence, sample observations are not I.I.D., aspect implying that the time series is nonlinearly dependent if the first differences of the natural logarithm are calculated.

According to Chu (2001), The BDS test statistic employs the concept of spatial correlation from the chaos theory. To compute this test statistic, the following procedures are followed:

1- Given the time series with *N* observations that should be the first difference of the natural logarithms for raw data in time series.

$$\{x_i\} = [x_1, x_2, x_3, \dots \dots, x_N] \qquad (\mathbf{1})$$

2- Choose a value of embedding dimension (*m*); embed a time series into *m*-dimensional vectors through taking each *m* consecutive points in the data series. This transforms the scalars series into vectors series with overlapping entries.

$$\begin{aligned} x_1^m &= (x_1, x_2, \dots, x_m) \\ x_2^m &= (x_2, x_3, \dots, x_{m+1}) \\ &\vdots \\ x_{N-m}^m &= (x_{N-m}, x_{N-m+1}, \dots, x_N) \end{aligned} \qquad (\mathbf{2})$$

3- Computing the correlation integral that measures the spatial correlation among the points through adding many pairs of points (*i, j*), where $1 \le i \le N$ and $1 \le j \le N$, in *m*-dimensional space, which are close in the sense that the points are within the radius or tolerance (ε) of each other.

$$C_{\varepsilon,m} = \frac{1}{N_m(N_m-1)} \sum_{i=j} I_{i,j;\varepsilon} \qquad (\mathbf{3})$$

Where, $I_{i,j;\varepsilon} = 1 \qquad if \ \left\| x_i^m - x_j^m \right\| \le \varepsilon$

$= 0 \qquad$ otherwise

4- Brock et al. (1987) and Brock et al. (1996) presented that if the time series is independently and identically distributed (I.I.D).

$$\mathrm{C}_{\varepsilon,\mathrm{m}} \approx \left[\mathrm{C}_{\varepsilon,1}\right]^{\mathrm{m}} \qquad (\mathbf{4})$$

and the statistic:

$$B_{\varepsilon,m} = N_T^{1/2}\left[C_{\varepsilon,m} - \left(C_{\varepsilon,1}\right)^m\right] \qquad (\mathbf{5})$$

If, $\frac{\mathrm{N}}{\mathrm{m}} > 200$, the values of ratio $\frac{\varepsilon}{\sigma}$ range from 0.5 to 2 (where σ is the standard deviation of the selected data), and the values of *m* are between 2 and 5, the quantity $\left[C_{\varepsilon,m} - \left(C_{\varepsilon,1}\right)^m\right]$ has an asymptotic normal distribution with 0 mean and a variance $V_{\varepsilon,m}$ can be stated as:

$$V_{\varepsilon,m} = 4\left[K^m + 2\sum_{j=1}^{m-1} k^{m-j} c_\varepsilon^{2j} + (m-1)^2 c_\varepsilon^{2m} - m^2 K c_\varepsilon^{2m-2}\right] \qquad (\mathbf{6})$$

Where, $k = k_\varepsilon = \frac{6}{N_m(N_m-1)(N_m-2)} \sum_{i<j<N} h_{i,j,N,\varepsilon}$;

$$h_{i,j,N,\varepsilon} = \frac{\left[I_{i,j,\varepsilon}I_{j,N,\varepsilon} + I_{i,N,\varepsilon}I_{N,j,\varepsilon} + I_{j,i,\varepsilon}I_{i,N,\varepsilon}\right]}{3}$$

Thus the statistic:

$$W_{\varepsilon,m} = B_{\varepsilon,m}/\sqrt{V_{\varepsilon,m}} \qquad (7)$$

Known as the BDS statistic, so the BDS test statistic can be given as:

$$BDS_{\varepsilon,m} = \frac{\sqrt{N}\left[C_{\varepsilon,m} - \left(C_{\varepsilon,1}\right)^{m}\right]}{\sqrt{V_{\varepsilon,m}}} \qquad (\mathbf{8})$$

BDS statistic converges to the standard normal distribution N(0,1) and statistical inference is possible.

The BDS test statistic is a two-tailed test, it is necessary to reject the null hypothesis if the BDS test is less than or greater than the critical values; for example, if α= 0.05, the critical value = ±1.96. Brock et al. (1991) demonstrated that the normal distribution was found to be a good approximation for the distribution of the BDS statistic when there were greater than 500 observations. One of the important points to be recognized that the BDS test statistic depends largely on the choice of values for ε and *m*. With a large (or small) ε the spatial correlation between the points of data will tend to be high (or low). The larger is the embedding dimension the smaller will be the number of non-overlapping histories, and as a result, the points selected by embedded vectors will become closer and the BDS statistic values will tend to be higher. According to the recommendation that suggests by Brock et al. (1991); Panagiotidis (2002); Marinela (2014), it is possible to set ε/σ = 0.5 to 2, and *m*= 2 to 4.

The researcher explained above how the BDS test can be used to test raw data; however, the BDS test can be employed on the filtered time series. In particular, the BDS test statistic on filtered time series is a particular application of the BDS test, where linear or nonlinear stochastic model is first fitted for the original data and the residual data is subject to the BDS test (see Karytinos (1999); Bhattacharya and Sensarma (2013); Bonilla et al. (2011)). Therefore, the BDS test on the residuals can be used to analyze whether the best fit model for a selected time series is a stochastic model or there is a potential deterministic nonlinearity in the data.

3.2.4.2. Empirical Evidence

First, the series has been filtered by fitting the best linear time series model to remove any linear dependence in the series. To choose the appropriate linear model, the partial autocorrelograms have been used in figure (2). These criteria refer to the proper AR lag for the linear model to be fitted and in the case that a distinct lag is referred by each method, the longer one is selected. In our study criteria indicate an AR (2) model specification for the EGX30 series.

Table 4 presents the BDS test statistic for the original "unfiltered" EGX30 series, the AR(2) residuals "filtered data", the scrambled residuals AR(2) "or randomly shuffled" residuals. The randomly shuffled residuals serve as an additional measure to control the

reliability of the BDS test because shuffling will destroy the assumed non-linearity in the data and the BDS statistic test will distinguish the difference between the shuffled and the unshuffled series.

As presented in the table, BDS test results strongly suggest that EGX30 returns series are nonlinearity dependent. The test statistic is significant at the confidence level of 1% (the critical value is 2.576 for the normal distribution) in each case of the original and filtered data under all embedding dimensions *m* and *r* sizes that were used, and therefore rejecting the IID-null hypothesis. Note that filtering data reduces the values of the W statistic value, which indicates that current linear dependence affects the BDS results. Consequently, it is clear that BDS statistics are non-significant (at the same confidence level of 1%) in both cases of the scrambled residuals Data and the Random Data "Gaussian white noise".

Table 4. The BDS test (W statistic) for the EGX 30 returns, Filtered returns, scrambled filtered returns and a Gaussian random surrogate

Length (in std. dev.) ε	Embedding dimension m	Original data	Filtered data (AR2 resid.)	Scrambled residuals	Random data
0.5	2	16.7397	16.7334	-0.0346	0.5024-
	3	23.4291	23.4717	0.1753	-0.2245
	4	28.8748	28.9245	0.4218	-0.5238
	5	36.0236	36	0.3467	0.6014
1	2	18.0141	17.9740	-0.0410	0.5594
	3	23.8744	23.8652	0.2499	-0.0124
	4	27.6767	27.7084	0.6766	-0.4179
	5	31.1724	31.2096	0.7020	0.5127
1.5	2	18.9026	18.8391	-.4057	0.8816
	3	23.7487	23.6847	0.0762	-0.4771
	4	26.4268	26.3686	0.2849	-0.0223
	5	28.4695	28.4212	.2790	0.1309
2	2	18.9866	18.8927	-0.8409	1.0932
	3	22.6443	22.5552	-0.1707	0.9292
	4	24.4065	24.3248	-0.1798	0.4399
	5	25.6250	25.5607	0.2162	0.3749

According to results are presented in Table 4, The BDS test (W statistic) is significant at the 1% level for all different parameters used. The results indicate the presence of nonlinearities in the EGX30 daily returns series, the previous results indicate only the presence of nonlinearity and a non-IID in returns series that can be of a stochastic or of a chaotic nature. Otherwise, it can be said that the results of the BDS test indicate that a chaotic interpretation cannot be ruled out for our series.

3.2.5 The Rescaled Range (R/S) Analysis

3.2.5.1. Description of Rescaled Range (R/S) Analysis

According to Sprott (2003); Peters (1994) and Karytinos (1999), rescaled range analysis (R/S) is a very useful tool in investigating the fractal structure[1] and non-periodic cycles in time series. Stability is banded and the presence of long-term correlations implies structure. The type of hidden structure detected from R/S analysis is consistent with chaotic processes.
Technically, the R/S method includes the time dimension by examining whether the range of fluctuations change, depending on the length of time used for measurement of its origins are related to the formula describing the Brownian Motion "B.M" (see Fan (2013); Peters (1994); Karytinos (1999); Krištoufek (2010)):

$$R = T^{0.5} \qquad (\mathbf{7})$$

where: R= the distance covered by a random particle suspended in a fluid, and T= a time index.

Equation (7) displays how R is scaling with time T in the case of a random system and this scaling is given by the slope of the log (R) vs. log (T) plot, which is equal to 0.5. Yet, when a system or a time series is not independent, namely not a random B.M., (7) cannot be applied, so, Hurst gave the following generalization, which can be used in this case:

$$(R/S)_n = cn^H \qquad (\mathbf{8})$$

where: $(R/S)_n$ = the rescaled range statistic measured over a time index n, c = a constant and H= the Hurst exponent, which displays how the R/S statistic is scaling with time.

The objective of the R/S method is to estimate the Hurst exponent (H), which, as shall be seen, can characterize a time series. This can be done by transforming (8) to:

$$\log(R/S)_n = \log(c) + H\log(n) \qquad (\mathbf{9})$$

Given a time series $\{X_t : t = 1, \dots N\}$, the R/S statistic can be defined as the range of cumulative deviations from the mean of the series, rescaled by the standard deviation. The analytical procedure to estimate the (R/S), values, as well as, the Hurst exponent by applying (9), is described in the following steps:

Step 1: The time period spanned by the time series of length N, is divided into p contiguous sub-periods of length n such that pn =N. The elements in each subperiod $X_{i,j}$, have two subscripts, the first (i=1,..., n) to denote the number of elements in each sub-periods and the second (j=1,..,p) to denote the sub-periods index. For each sub period j, the R/S statistic is calculated as follows:

(1) Fractal structure is symptomatic of nonlinear feedback effects, resulting in repetitive structure in smaller and smaller scales.

$$\left(\frac{R}{S}\right)_j = S_j^{-1}\left[\max_{1\le K\le n}\sum_{i=1}^{K}\left(X_{ij}-\bar{X}_j\right) - \min_{1\le K\le n}\sum_{j=1}^{K}\left(X_{ij}-\bar{X}_j\right)\right] \qquad (\mathbf{10})$$

where: s_j is the standard deviation for each subperiod.

Step 2: The $(R/S)_n$, namely the R/S statistic for time length n, is given by the average of the $(R/S)_j$ values for all the p contiguous subperiods with length n, as:

$$\left(\frac{R}{S}\right)_n = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{R}{S}\right)_j \qquad (\mathbf{11})$$

Step 3: Equation (11) gives the R/S value, which corresponds to a certain time interval of length n. To apply equation (9), steps one and two are repeated by increasing n to the next integer value, until n = N/2, since, at least two subperiods are needed, to avoid bias.
From the above procedure, the Hurst exponent can be estimated, as follows (see Sprott (2003); Peters (1994); Karytinos (1999); Davidsson (2011))**:**

1- The Hurst exponent takes values from 0 to 1 ($0 \le H \le 1$). Gaussian random walks, or, more generally, independent processes, give H=0.5.
2- If $0.5 \le H \le 1$, positive dependence is indicated, and the series is called persistent or trend reinforcing. In this case, the series is characterized by a long memory process with no characteristic time scale. The lack of a characteristic time scale and the existence of power law (the log/log plot) are the key characteristics of a fractal series.
3- If $0 \le H \le 0.5$, negative dependence is indicated, yielding anti-persistent behavior (only if the system under study is assumed to have a stable mean).

However, chaotic systems have also Hurst exponents H= 0.5, and in chaotic terms, long memory effects correspond to Sensitive Dependence on Initial Conditions (SDIC), the SDIC property combined with fractality characterizes chaotic systems. Pure chaotic processes have Hurst exponents close to 1.
When dealing with real data, the problem of distinguishing between the above-mentioned alternatives becomes more difficult due to the existence of noise. Most series are contaminated by either additive or dynamical (system) noise. Hence, in most cases, which include financial data, the problem is to distinguish between fractal noise and noisy chaos. On the other hand, R/S analysis provides a very useful tool to solve the above problem, that is able to detect, even when noise is present, the existence of cycles in the series and thus to characterize it.
According to Karytinos (1999), there are many ways to detect cycles by the use of R/S analysis and estimate their length. However, the researcher will suffice in the present study with discussing two methods only; they are:
1- The regressions of log (R/S) vs. log (n) can be used for different n-lengths. The n value, for which a peak value of H is obtained, corresponds to the cycle length.

2- Finally, a second and more efficient way to find the cycle length is given by the V-statistic Peters (1994); McKenzie (2001) defined as:

$$V_n = (R/S)_n / n^{0.5} \qquad (12)$$

The V_n vs. log (n) plot gives a flat line for an independent random process and an upwardly sloping curve in the case of persistent series. The existence of a cycle and its length can be discerned (even when noise is present) from the break-point in this plot occurring when V_n reaches a peak and then flattens out, an indication that the long-memory process has dissipated.

3.2.5.2. Empirical Evidence

As mentioned before, the EGX 30 returns consist of 4393 daily observations covering 17 years. Since the series exhibits serial autocorrelation the AR (2) filtered series have been also used in rescaled range (R/S) analysis, as presented in table 5.

Table 5. Results of the R\S analysis for daily EGX30 returns

Sub-period length (n)	Log (*n*)	Original series		Filtered series		Scrambled residuals		Random data	
		Log R/S	*V*-statistic R/S	Log R/S	*V*-statistic R/S	Log R/S	*V*-statistic R/S	Log R/S	*V*-statistic R/S
10	1	0.4968	0.9926	0.4945	0.9873	0.4506	0.8925	0.4694	0.932
14	1.1461	0.6026	1.0705	0.5953	1.0525	0.5396	0.9259	0.5818	1.0203
20	1.3010	0.7109	1.1492	0.7056	1.1352	0.6259	0.9448	0.6694	1.0445
25	1.3979	0.769	1.1751	0.7709	1.18	0.7072	1.0191	0.7053	1.0146
28	1.4472	0.8079	1.2142	0.8079	1.2144	0.74	1.0386	0.7377	1.0331
35	1.5441	0.865	1.2386	0.8633	1.2339	0.7763	1.0098	0.7781	1.0141
40	1.6021	0.9034	1.2659	0.9076	1.278	0.8018	1.0017	0.8414	1.0974
50	1.6990	0.9672	1.3114	0.9712	1.3234	0.8786	1.0695	0.9148	1.1622
56	1.7482	1.0155	1.3847	1.014	1.3801	0.8925	1.0433	0.9203	1.1122
70	1.8451	1.0518	1.3468	1.0345	1.2941	0.9458	1.0551	0.9729	1.1229
100	2.0000	1.1263	1.3375	1.111	1.2913	1.0403	1.0973	1.0791	1.1999
125	2.0969	1.1817	1.3591	1.1668	1.3133	1.0917	1.1046	1.1396	1.2336
140	2.1461	1.1717	1.2548	1.1625	1.2287	1.1277	1.134	1.1422	1.1726
175	2.2430	1.2372	1.3053	1.2257	1.2711	1.1639	1.1024	1.1759	1.1335
200	2.3010	1.3008	1.4135	1.2947	1.3938	1.196	1.1104	1.2358	1.217
250	2.3979	1.3616	1.4544	1.3396	1.3825	1.2978	1.2556	1.3327	1.3606
280	2.4472	1.3815	1.4386	1.379	1.4304	1.277	1.131	1.3417	1.3127
350	2.5441	1.4521	1.5137	1.4303	1.4397	1.3687	1.2494	1.4087	1.3698
500	2.6990	1.5243	1.4956	1.5195	1.479	1.4116	1.1538	1.5401	1.551
700	2.8451	1.6995	1.8923	1.6935	1.8664	1.4527	1.0719	1.4476	1.0593
875	2.9420	1.7117	1.7407	1.7085	1.7279	1.558	1.2217	1.5641	1.2392
1000	3.0000	1.7326	1.7084	1.726	1.6827	1.5295	1.0703	1.6687	1.4745
1240	3.0934	1.741	1.5643	1.7314	1.5299	1.5905	1.106	1.7511	1.6009

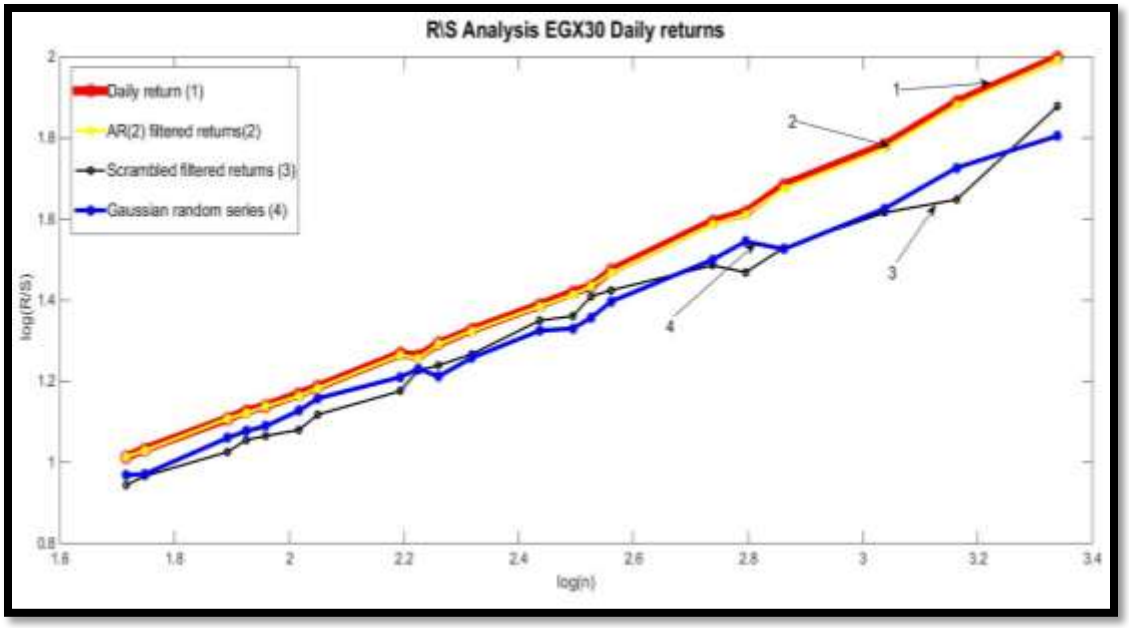


Fig. (5): Log (R/S) vs. Log (n) plot for the H exponent estimation

Figures 5 and 6 present the log-log plot of the R/S statistic versus time and the semi-log plot of the V-statistic respectively for four different series. The original returns, the filtered returns, a random IID series of scrambled filtered returns and a random Gaussian version of the filtered returns.

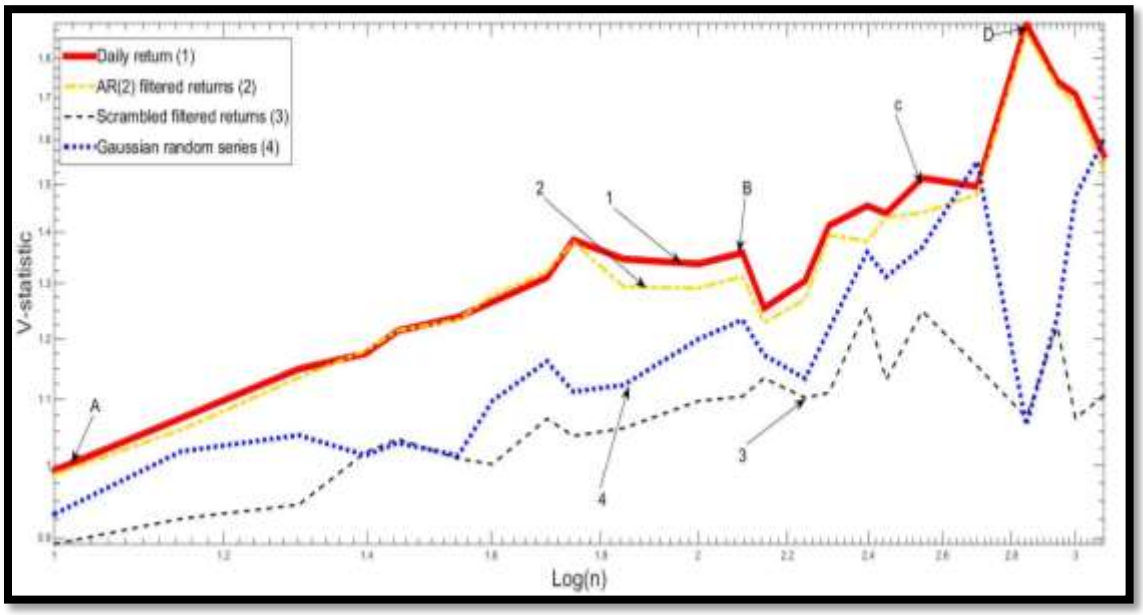


Fig. (6): V-statistic vs. Log (n) plot for the cycle-length estimation

The V-statistic plot demonstrates a possible long-term cycle with a length of 1240 days for both the original and filtered series, while for the scrambled residuals and the Gaussian surrogate a much flatter curve occurs. However, a smaller cycle appears also, having a length of approximately 125 days as can be seen in Figure 6, the V-statistic

curve for the raw and filtered series rises up from point A to point B corresponding to 125 days, then from B to C "approx. 350 days" it flattens out to rise up again to point D "1240 days".

To verify these findings, the Hurst exponent has been calculated for each cycle, as well as for the intermediate between the two cycles period and bootstrapping has been used to assess the statistical significance of the H estimates against both the Gaussian random and the random IID alternatives.

Table 6. Hurst estimates test of the daily EGX30 returns

		Time period (in trading days)			
Part A		(A to B) 1<n<125	(B to C) 125 < n<350	(C to D) 351<n<700	(A to D) 1<n<1240
	H estimate (Original series)	0.6075	0.6856	0.4705	0.5043
	H estimate (Filtered series)	0.6065	0.6594	0.4156	0.5036
Bootstrapping results					
Part B	**H estimate (Scrambled Residuals)** H_0 **:H =** H_S	0.5727	0.6762	0.7038	0.4466
	significance	0.0359	0.0116	0.0149	0.0102
	H estimate (Random Data) H_0 **:H =** H_R	0.5621	0.5447	0.4094	0.5448
	Significance	0.0488	0.0178	0.0110	0.0041

In Table 6, capital letters in parentheses beside each column number corresponding to the time periods signified in the V-statistic plot.

Part A of Table 6 presents the R/S analysis findings. It can be seen that the H estimates from the filtered series for all the different periods tested are lower than the ones from the original data, due to the linear autocorrelation bias that has been removed.

Part B presents the bootstrapping results of testing the H estimate from the filtered series, for each time period, against the two random alternatives. In Table 6, it is possible to see the mean H estimate from 5000 bootstrapped samples corresponding to each random alternative and below it, the significance level that defines whether the random null hypothesis is rejected or not.

However, the H estimates of 0, 6065 for the smaller (125 days) cycle and of 0, 5036 for the longer (1240 days) cycle, are quite low, reflecting the existence of a strong noisy component in the series.

Table 7. The modified R\S statistic of the daily EGX30 returns

Q	Andrew's 12	$N^{1/4}$ 8	$N^{1/3}$ 16	$N^{1/2}$ 66	100	150
V_q – statistic	1.9079	1.9269	1.8764	1.6803	1.5520	1.4556
Bootstrapped						
Q	**Andrew's 12**	$N^{1/4}$ **8**	$N^{1/3}$ **16**	$N^{1/2}$ **66**	**100**	**150**
V_q – statistic	1.2390	1.2498	1.2304	1.2725	1.2682	1.2404

Significance at 0.5% level according to the asymptotic critical values (.05=0.721)

It is possible to see from Table 7 that for all the q values employed, the V_q estimates reject the test's short-range dependence null when compared to either the asymptotic or the bootstrapped critical values of the test.

The results from the modified R/S statistic seem to confirm the above findings concerning the fractality and long-term dependence of the EGX30 series.

4. Conclusion

The limitations and inadequacies of the traditional linear stochastic framework with respect to explaining the dynamics of the behavior of financial markets have led the recent research to new nonlinear approaches. Among them, chaotic models have attracted increasing attention since they have been demonstrated to exhibit interesting theoretical and empirical features that could help to gain a better insight into the underlying financial markets mechanisms.

In empirical terms, this research focuses on the financial markets: the Egyptian stock exchange, considered to be an emerging market. The researcher aimed at seeing whether the movement of the EGX30 prices index can be explained through chaotic models. Some steps have been taken to verify this, they are as follows:

- The testing of the EGX30 series starts with statistical analysis in the time and frequency domains, where the series is found to exhibit serial autocorrelation and deviations from normality with fat tails in their distributions.
- In the next step, the BDS test, a powerful test for independence, which after proper treatment of the data can also be used as a test for nonlinearity, indicates that the series has nonlinear components. This result was compatible with a chaotic explanation, so the researcher proceeded further with the R/S test.

The latter can reveal long-term memory and fractal structure of the series analyzed, even in the presence of noise. Moreover, it can distinguish between fractal noise processes and noisy chaos specifications and the R/S test has been applied with the bootstrap method for safer and statistically based conclusions and gave the first indications of different structural characteristics in our series, where the EGX30 series was found to be noisy, fractal and persistent with (1240 days) cycle.

Therefore, the result is according to the above, a noisy chaos explanation was favored against a fractal noise one for the EGX30 returns index.

References

Afonso, A., & Teixeira, J. (1998). Non-linear tests of weakly efficient markets: Evidence from Portugal.

Aguirre, L. A., & Letellier, C. (2009). Modeling nonlinear dynamics and chaos: a review. Mathematical Problems in Engineering, 2009.

Barton, S. (1994). Chaos, self-organization, and psychology. American Psychologist, 49(1), 5.

Baumol, W. J., & Quandt, R. E. (1985). Chaos models and their implications for forecasting. Eastern Economic Journal, 11(1), 3-15.

Bhattacharya, A., & Sensarma, R. (2013). Non-linearities in emerging financial markets: evidence from India. IIM Kozhikode Society & Management Review, 2(2), 165-175.

Bonilla, C. A., Romero, R., & Gutierrez, E. (2011). Episodic Non-Linearities and Market Efficiency in the Mexican Stock Market. The Manchester School, 79(3), 367-380

Brock, W. A., Hsieh, D. A., LeBaron, B. D., & Brock, W. E. (1991). Nonlinear dynamics, chaos, and instability: statistical theory and economic evidence. MIT press.

Brock, W. A., Scheinkman, J. A., Dechert, W. D., & LeBaron, B. (1996). A test for independence based on the correlation dimension. Econometric reviews, 15(3), 197-235.

Brock, W., Dechert, W. D., & Scheinkman, J. (1987). A test for independence based on the correlation dimension, University of Wisconsin. Economics Working Paper SSRI-8702.

Ch Mandelbrot, B. B. (1982). The many faces of scaling: fractals, geometry of nature, and economics. Self-Organization and Dissipative Structures: Applications in the Physical and Social Sciences Eds WC Schieve, PM Allen (University of Texas Press, Austin, TX) pp, 91-109.aos and Nonlinear Dynamics in the Egyptian Stock Market

Chandra, A., & Thenmozhi, M. (2017). Behavioural asset pricing: Review and synthesis. Journal of Interdisciplinary Economics, 29(1), 1-31.

Chu, H. G. A. C. K. (2001). Using BDS statistics to detect nonlinearity in time series.

Davidsson, M. (2011). Serial dependence and rescaled range analysis. International Research Journal of Finance and Economics, (64), 186-197.

Faggini, M., & Parziale, A. (2012). The failure of economic theory. Lessons from chaos theory. Modern Economy, 3(01), 1.

Fan, J. (2013). Rescaled range analysis in higher dimensions. Research Journal of Applied Sciences, Engineering and Technology, 5(18), 4489-4492.

Feigenbaum, M. J. (1978). Quantitative universality for a class of nonlinear transformations. Journal of statistical physics, 19(1), 25-52.

Gleick, J. (1987). Chaos: Making a New Science, Viking Penguin. Inc., New York.

Gunaratne, S. A. (2003). Thank you Newton, welcome Prigogine: Unthinking'old paradigms and embracing new directions. Part 1: Theoretical distinctions. Communications-Sankt Augustin Then Berlin, 28(4), 435-456.

Harriff, R. B. (1997). Chaos and nonlinear dynamics in the financial markets: Theory, evidence, and applications: Robert R. Trippi (ed.), Irwin Professional Publishing Company, Burr Ridge, IL, 1996, 528 pages, $85, ISBN 1-55738-857-7. International Journal of Forecasting, 13(1), 146-147.

Karytinos, A. D. (1999). A chaos theory and nonlinear dynamics approach to the analysis of financial series: a comparative study of Athens and London stock markets (Doctoral dissertation, University of Warwick).

Krištoufek, L. (2010). Rescaled range analysis and detrended fluctuation analysis: Finite sample properties and confidence intervals. Czech Economic Review, 4(3), 315-329.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. Journal of the atmospheric sciences, 20(2), 130-141.
Marinela, R. C. V. (2014). Complex Nonlinear Dynamics of Financial Time Series. An Empirical Analysis Using Chaos Theory (Doctoral dissertation, Master's dissertation, The Academy of Economic Studies, Bucharest, Romania).
May, R. M. (1975). Biological populations obeying difference equations: stable points, stable cycles, and chaos. Journal of Theoretical Biology, 51(2), 511-524.
McKenzie, M. D. (2001). Non-periodic Australian Stock Market Cycles: Evidence from Rescaled Range Analysis. Economic Record, 77(239), 393-406.
Oestreicher, C. (2007). A history of chaos theory. Dialogues in clinical neuroscience, 9(3), 279.
Panagiotidis, T. (2002). Testing the assumption of linearity. Economics Bulletin, 3(29), 1-9.
Peters, E. E. (1994). Fractal market analysis: applying chaos theory to investment and economics (Vol. 24). John Wiley & Sons.
Selvam, A. M. (2012). Nonlinear dynamics and chaos: Applications in atmospheric sciences. Journal of Advanced Mathematics and Applications, 1(2), 181-205.
Smale, S. (1963). Stable manifolds for differential equations and diffeomorphisms. Annali della Scuola Normale Superiore di Pisa-Classe di Scienze, 17(1-2), 97-116.
Sprott, J. C., (2003). Chaos and time-series analysis (Vol. 69). Oxford: Oxford University Press.

New Shrinkage Parameters for Liu-type Zero Inflated Negative Binomial Estimator

El-Housainy A. Rady[1], Mohamed R. Abonazel[1], & Ibrahim M. Taha[2*]

Abstract

Multicollinearity is a very common problem especially in applied sciences. It may lead to imprecision of the maximum likelihood (ML) estimates, resulting in uncertainty when assessing the effects of the explanatory variables on the response variable. Some biased estimators have been proposed in the literature for handling this dilemma. As a remedy to this problem in the area of zero inflated count models, Rady et al. (2019) proposed a Liu-type estimator for zero inflated negative binomial (ZINB) model as a general biased estimator which includes other biased estimators as special cases. This paper presents some shrinkage parameters for Liu-type zero inflated negative binomial (LTZINB) estimator. We compare the performance of the Liu-type estimator along with the maximum likelihood estimator through the mean squared error (MSE) and mean absolute error (MAE) by conducting a Monte Carlo simulation study. According to results; the Liu-type estimator outperforms the ML estimator in terms of MSE and MAE.

Key words: EM Algorithm, Liu-type Estimator, ML Estimator, Monte Carlo Simulation, Multicollinearity, ZINB.

1. Introduction

For count data, the standard framework for explaining the relationship between the response variable and a set of explanatory variables includes the Poisson and negative binomial regression models. However, the basic Poisson regression model forces the conditional variance of the outcome to equal the conditional mean, which is of limited

[1] Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt.
[2] Department of Mathematics, Statistics, and Insurance, Sadat Academy for Management Sciences, Tanta Branch, Egypt. *E-mail : ibrahimaboalazm@gmail.com

use in real life. The negative binomial regression can be written as an extension of Poisson regression and it enables the model to have greater flexibility in modeling the relationship between the conditional variance and the conditional mean compared to the Poisson model. Also, an often encountered characteristic of count data, is the number of zeros in the sample can exceed the number of zeros predicted by either Poisson or negative binomial model, and this is of interest because zero counts frequently have special status. In such a circumstance, the ZINB regression model better accounts for these characteristics compared to other count models.

It is well-known that the covariance matrix of ML estimate become ill-conditioned in the presence of sever multicollinearity, and as a result the variance of the regression coefficients gets inflated. Biased estimation has been a widely accepted technique in dealing with non-orthogonal problem in regression. Hoerl and Kennard (1970a,b) suggested the ridge estimator to handle this problem. The idea behind ridge estimator is to add a small positive amount (k) to the diagonal entries of the covariance matrix to improve the conditioning of this matrix, reduce the MSE, and obtain stable estimated coefficients. But there is a disadvantage of the ridge estimator that is the estimated parameters are nonlinear functions of the ridge parameter.

Liu (1993) introduced Liu estimator which is a linear function of the shrinkage parameter, it is a combination of the Stein estimator proposed by Stein (1956) and the ridge estimator. Latterly, Liu (2003) showed the superiority of the Liu-type estimator over the ridge estimator. Liu-type estimator is a two-parameter estimator that uses advantages of both ridge estimator and Liu estimator. This paper is organized as follows; section 2 presents a background about ZINB model and ML estimation using EM algorithm as a numerical technique method. Section 3 discusses the Liu-type estimator for ZINB model. A Monte Carlo simulation study in section 4 is conducted to investigate the performance of the LTZINB estimator with the proposed parameters. Finally, section 5 presents the results, conclusion, and future works.

2. Zero-Inflated Negative Binomial Model

Zero inflated negative binomial distribution introduced by Greene (1994) is formed of Poisson Gamma mixture distribution, with a dispersion parameter that useful to describe the variation of the data. It assumes that there are two distinct data generation processes, the result of a Bernoulli trial is used to determine which of the two processes is used. For observation i, with probability φ_i the only possible response of the first process is zero counts, and with probability of $(1-\varphi_i)$ the response of the second process is governed by a negative binomial with mean φ_i.

The zero counts are generated from the first and second processes, where the probability is estimated for whether zero counts are from the first or the second process. Let y_i be a nonnegative response count variable. Thus, the ZINB distribution can be defined as:

$$\mathrm{f}(y_i;\varphi_i,\mu_i,\theta)=\begin{cases}\varphi_i(1-\varphi_i)\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta} & y_i=0;\\ (1-\varphi_i)\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i} & y_i>0,\end{cases} \quad (1)$$

where $0\le\varphi_i\le 1, \mu_i \ge 0$, with $\theta>0$ and $\Gamma(.)$ is the gamma function. The mean and variance for the ZINB distribution are given by:

$$\mathrm{E}(y)=(1-\varphi_i)\mu;\ \mathrm{Var}(y)=\mu(1-\varphi_i)(1+\mu(\varphi_i+\theta)).$$

The ZINB for the response y_i can be written as:

$$y_i\sim\begin{cases}0 & \text{with probability } \varphi_{i_i};\\ \mathrm{NB}(y_i|\,\mu_i,\theta) & \text{with probability } 1-\varphi_{i_i},\end{cases} \quad (2)$$

with

$$\varphi_i=\frac{\exp(z_i'\gamma)}{1+\exp(z_i'\gamma)},$$

where z_i is the i^{th} row of Z which is the matrix for the logit model, and γ corresponds to a $(p+1)\times 1$ vector of coefficients. Furthermore, $\mu_i=\exp(x_i'\beta)$ where x_i is the i^{th} row

of X, which is the matrix for the ZINB model, and β corresponds to a $(p+1)\times 1$ vector of coefficients. The model allows μ_i and φ_i to depend on covariates through the relationships:

$$\mathrm{g}(\varphi_i) = \log\left(\frac{\varphi_i}{1-\varphi_i}\right) = z_i'\gamma;\ \mathrm{h}(\mu_i) = \log\left(\frac{\theta\mu_i}{1+\theta\mu_i}\right) = x_i'\beta.$$

The regression coefficients can be estimated using the ML method. The log-likelihood function for the parameters (γ, β, θ) and γ is given by:

$$\mathcal{L}(\gamma,\theta,\beta;Y) = \sum_{y_i=0}\log\left\{\varphi_i(1-\varphi_i)\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\right\}$$
$$+\sum_{y_i>0}\log\left[(1-\varphi_i)\left(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i}\right)\right].$$

Let an indicative function D_i defined as follows:

$$D_i = \begin{cases} 1 & \text{if} \quad y_i = 0; \\ 0 & \text{otherwise}. \end{cases}$$

The log-likelihood can be rewritten as follows:

$$\mathcal{L}(\gamma,\theta,\beta;Y) = \sum_{i=1}^{n} D_i \log\left\{\varphi_i(1-\varphi_i)\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\right\}$$
$$+\sum_{i=1}^{n}(1-D_i)\log\left[(1-\varphi_i)\left(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i}\right)\right]. \qquad (3)$$

To obtain the parameter estimates of ZINB regression model, we can use the expectation–maximization (EM) algorithm to maximize the log-likelihood function, the reason behind choosing the EM algorithm is that it's straightforward to implement since the iterations converge monotonically and thus guarantee a good stability of the estimation procedure. Moreover, the EM estimates are quite insensitive to the initial values; See Dempster et al. (1977), Lambert (1992), and Nanjundan (2006). This algorithm takes the conditional expectation of the log-likelihood (the E-step), maximizes it with respect to the required parameters (the M-step), then alternates between the two steps (the E-step and the M-step) until convergence.

The missing data are defined as indicator variables $\delta = (\delta_1, \delta_2, \dots, \delta_n)'$ where $\delta_i = 1$ when y_i is from the zero case, and δ_i = 0 when y_i is from the count case. The complete-data log-likelihood can be defined as follows:

$$\mathcal{L}\,(\beta, \gamma, Y, \delta) = \mathcal{L}_1(\gamma, y, \delta) + \mathcal{L}_2(\beta, y, \delta),$$

where

$$\mathcal{L}_1(\gamma, Y, \delta) = \sum_{i=1}^{n}[\delta_i z_i y - \log(1 + \exp(z_i'\gamma))]\,,$$

and

$$\mathcal{L}_2(\beta, Y, \delta) = \sum_{i=1}^{n}(1 - \delta_i)\log\left(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\exp(x_i'\beta)}\right)^{\theta}\left(\frac{\exp(x_i'\beta)}{\theta+\exp(x_i'\beta)}\right)^{y_i}\right).$$

This log-likelihood function can be easily maximized through the following steps:

Step (1): (Initial Value)

Specify initial values $\beta^{(0)}$ and $\gamma^{(0)}$.

Step (2): (E step)

Estimate δ_i for the Negative Binomial model by its conditional expectation $\delta_i^{(r)}$given by the current estimates $\beta^{(r)}$ and $\gamma^{(r)}$

$$\delta_i^{(r)} = \begin{cases} \left[1 + \exp\left(-z_i'\,\gamma^{(r)}\right)\left(\frac{\theta^{(r)}}{\exp(x_i'\beta^{(r)})+\theta^{(r)}}\right)^{\theta^{(r)}}\right]^{-1} & \text{if} \quad y_i = 0; \\ 0 & \text{if} \quad y_i > 0. \end{cases}$$

Step (3): (M step)

Finding $\gamma^{(r+1)}$ by maximizing $\mathcal{L}_1(\gamma, Y, \delta)$, this can be done by fitting an unweighted logistic regression of the response variable Y on the covariate matrix z with weights $1 - \delta^r$. And also finding $\beta^{(r+1)}$ by maximizing $\mathcal{L}_2(\beta, y, \delta^r)$ by fitting a weighted count regression of Y on the covariate matrix x with weights $1 - \delta^r$.

Step (4): (Stopping Rule)

Set $\beta^{(0)} = \beta^{(r)}$ and $\gamma^{(0)} = \gamma^{(r)}$ and continue from Step 2 until convergence occurs.

3. Zero Inflated Negative Binomial Liu-type Estimator

Little works have been done in the context of handling multicollinearity in zero inflated count models (see; Kibria et al. (2013), Yüzbaşi and Asar (2018), Rady et al. (2018), and Taha (2019)). The general formula for the ridge and Liu estimators are respectively as follows:

$$\hat{\beta}_{\mathrm{R}} = (X'\widehat{W}X + kI_p)^{-1}(X'\widehat{W}X)\hat{\beta}_{\mathrm{ML}}\ ;\ k > 0, \tag{4}$$

$$\hat{\beta}_{\mathrm{d}} = (X'\widehat{W}X + I_p)^{-1}(X'\widehat{W}X + dI_p)\hat{\beta}_{\mathrm{ML}}; 0 < d < 1, \tag{5}$$

where $\hat{\beta}_{ML} = (X'\widehat{W}X)^{-1}X'\widehat{W}Y;\ \widehat{W} = \mathrm{diag}(\hat{\mu}_i)$. The Liu-type estimation of the count regression models is introduced by Huang and Yang (2014), and Asar and Genç (2018). Recently, Rady, et al (2019) proposed a Liu-type estimator for ZINB model based on EM estimates as:

$$\hat{\beta}_{LT} = (X'\widehat{W}X + kI)^{-1}\big(X'\widehat{W}X + kdI\big)\hat{\beta}_{EM};\ k > 0, 0 < d < 1, \tag{6}$$

where $\hat{\beta}_{EM}$ is the vector of estimated coefficients using EM algorithm. The matrix mean squared error (MMSE) of the Liu-type estimator equals:

$$\mathrm{MMSE}\big(\hat{\beta}_{LT}\big) = (X'\widehat{W}X + kI)^{-1}\big(X'\widehat{W}X + kdI\big)(X'\widehat{W}X)^{-1}\big(X'\widehat{W}X + kdI\big)(X'\widehat{W}X + kI)^{-1} + k^2(1-d)^2(X'\widehat{W}X + kI)^{-1}\alpha\alpha'(X'\widehat{W}X + kI)^{-1}, \tag{7}$$

and the total mean squared error (TMSE) of it in canonical form:

$$\mathrm{TMSE}\big(\hat{\beta}_{LT}\big) = \sum_{j=1}^{p}\left\{\frac{(\lambda_j + kd)^2}{\lambda_j(\lambda_j+k)^2} + \frac{k^2(1-d)^2\alpha_j^2}{(\lambda_j+k)^2}\right\}, \tag{8}$$

where λ_j are the ordered eigenvalues of $(X'\widehat{W}X)$, and $\alpha = \big(\alpha_1, \alpha_2, \dots, \alpha_p\big)' = V\beta$; where V is an orthogonal matrix whose columns are the corresponding eigenvectors of $\lambda_1, \lambda_2, \dots, \lambda_p$. The optimal value of the shrinkage parameter d can be obtained by differentiating equation (8) with respect to d, equating to zero, and solving for d as:

$$d_{opt} = \sum_{j=1}^{p}\frac{k(k\alpha_j^2 - 1)}{(\lambda_j+k)^2}\Big/\sum_{j=1}^{p}\frac{k(1+\lambda_j\alpha_j^2)}{\lambda_j(\lambda_j+k)^2}.$$

For more details about calculating d_{opt}, see Abonazel and Farghali (2019). The shrinkage parameter for LTZINB estimator will be the d optimal along with four ridge parameters chosen from the literature, resulting in four new parameters as shown in table (1).

Table (1): Shrinkage Parameters

Shrinkage Parameter		Author
d_{opt}	$K1 = \hat{\sigma}^2/\hat{\alpha}^2_{max}$,	developed by Hoerl et al. (1975),
	K3= $\hat{\sigma}^2/\left(\prod_{j=1}^{p}\hat{\alpha}_j^2\right)^{\frac{1}{p}}$,	developed by Kibria (2003),
	K5= $Max\left(1/m_j\right)$	developed by Muniz and Kibria (2009),
	K10= $median(m_j)$	developed by Kibria et al (2011).

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2}{n-p-1}$, $\hat{\alpha}^2_{\max} =$ is the maximum element of $\hat{\alpha}_j^2$, and $m_j = \sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_j^2}}$. See appendix (A) for full list of shrinkage parameters used in the area of generalized linear models.

4. Monte Carlo Simulation

In this section, a simulation study has been conducted to compare the performances of ML estimator and the suggested Liu-type estimators.

4.1 Design of Simulation

According to Alkhamisi and Shukur (2008), Zeebari et al (2012), Abonazel and Farghali (2019) and Abonazel (2018), the explanatory variables were generated from a multivariate normal distribution $MVN(1, \Sigma_X)$, where variance-covariance matrix Σ_X is defined as $\text{diag}(\Sigma_X) = 1$ and $\text{off} - \text{diag}\ (\Sigma_X) = \rho$. While the parameters are chosen to be $\sum_{j=1}^{p}\beta_j^2 = 1$ and $\beta_1 = \beta_2 = \cdots = \beta_p$, which are common restrictions in many simulation studies such as Månsson and Shukur (2011) and KaÇiranlar and Dawoud (2018).

The dependent variable has two distinct data generation processes the first one is binary variable from binomial distribution and the second one is the zero-inflated

distribution which is generated through equation (2). In this study, we generate intercept-logit model with regression coefficient $\gamma = 1$, then $\pi_i = \varphi_i = 0.5$. While the dispersion parameter $\theta = 2$ in generated ZINB model. In our simulation study, the effective factors are chosen to be the number of explanatory variables ($p = 2, 4, 6$), the sample size ($n = 100, 200, 300, 400$), and the correlation among the explanatory variables ($\rho = 0.7, 0.8, 0.9, 0.95$). The estimated MSEs and MAEs of the estimators to be examined are calculated as follows:

$$\mathrm{MSE}(\hat{\beta}) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\beta}_l - \beta)'(\hat{\beta}_l - \beta);\ \mathrm{MAE}(\hat{\beta}) = \frac{1}{L}\sum_{l=1}^{L}\left|\hat{\beta}_l - \beta\right|, \quad (9)$$

where $\hat{\beta}_l$ is the vector of estimated values at l^{th} experiment of $L = 2000$ Monte Carlo experiments, while β is the vector of true parameters. The programs to set up the Monte Carlo simulation study are written in R language.

4.2 Results of Simulation

Results of simulation are given in tables B.1, B2, and B.3 in Appendix B. The factors affecting the MSE and the MAE values of the estimators in the simulation are the degree of correlation ρ, the number of explanatory variables p, and the sample size n. The MSEs and the MAEs of MLE, and Liu-type estimators with the optimal shrinkage parameter are given. According to these tables, the values of MSE and MAE are increasing when both the number of explanatory variables and correlation between the independent variables increase. On the contrary the values of MSE and MAE are decreasing when the sample size increases. We can also observe that the proposed estimators have lower MSE and MAE compared to other parameters. As a result we can conclude that the proposed estimator LT(K_5) has better performance than the other estimators for different factors used in the simulation study.

4.3 Mean Relative Efficiency

Another comparative performance called relative efficiency (RE) can be used, it is calculated based MSE and MAE as follows:

$$\text{RE}\left(LT(k1)\right) = \frac{MSE\left(LT(k1)\right)}{MSE\ (ML)},\ \text{RE}\left(LT(k3)\right) = \frac{MSE\left(LT(k3)\right)}{MSE\ (ML)},$$
$$\text{RE}\left(LT(k10)\right) = \frac{MSE\left(LT(k10)\right)}{MSE\ (ML)}, \text{ and RE}\left(LT(k5)\right) = \frac{MSE\left(LT(k5)\right)}{MSE\ (ML)}. \quad (10)$$

The RE based MAE is calculated by the same way as the RE based MSE in equation (10). We report the results of Average REs (MREs) in a form of graphs in fig. 1 to fig. 3. These graphs are plotted using Microsoft Power BI. Fig. 1 is concerned with the MRE based MSE for LTZINB estimators, Fig. 2 for the MRE based MAE for LTZINB estimators, and Fig. 3 for the MRE based MSE and MAE for LTZINB estimators. It can be observed from the graphs, that in all models whenever the degree of correlation (rho) increases, the RE ratio decreases. On the other hand, in all cases the ratio of RE increases when the sample size increases except for $LT(k5)$.

It can also be concluded that, there are some obvious deviations in the REs between the selected estimators in all cases except when p=6 where the performance of $LT(k1)$ and $LT(k3)$ estimators are somehow identical. For all graphs and all cases we can observe that $LT(k5)$ has lower MRE than the other estimators, and as a result we can reach to the conclusion that $LT(k5)$ is more efficient and reliable than the other estimators. We may also notice from fig. 3.a that the MSE values are less than the MAE values. Fig. 3.b shows that the MRE for the simulated model with 6 independent variables is lower than that of 2 independent variables. Fig. 3.c investigate the different degree of correlations on the performance of the selected estimators, one may see that as the degree of correlation increases, the MRE decreases. Finally in fig. 3.d, it can be observed that as the sample size increases, the ratio of RE increases except for $LT(k5)$ the ratio decreases.

The Final conclusion from the simulation study is that the Liu-type estimator outperforms the ML estimator in the sense of MSE and MAE criteria. Moreover, $LT(k5)$ has the best performance in the simulation in all cases.

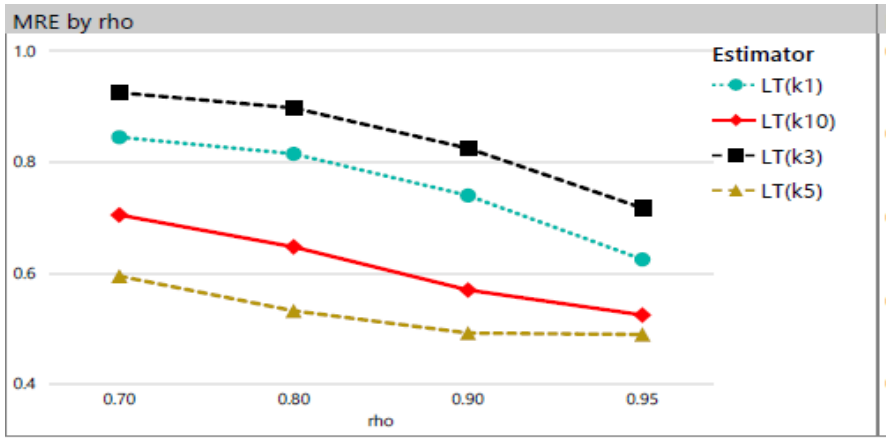


Fig. 1.a: MRE based MSE categorized by degree of correlation when p=2

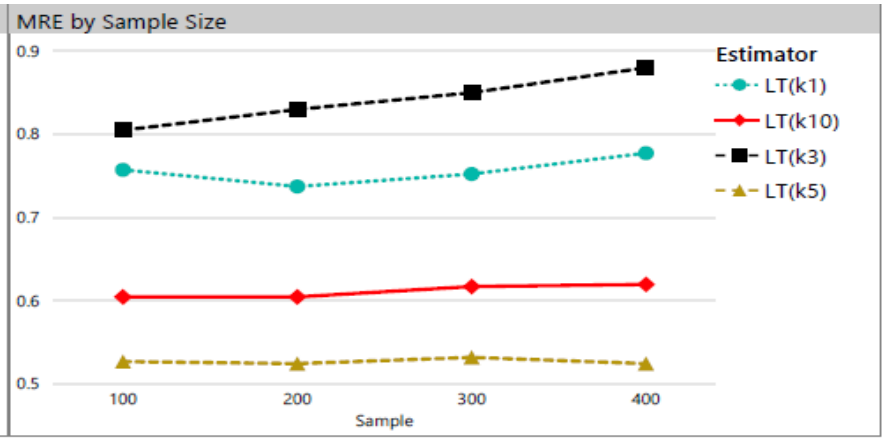


Fig. 1.b: MRE based MSE categorized by sample size when p=2

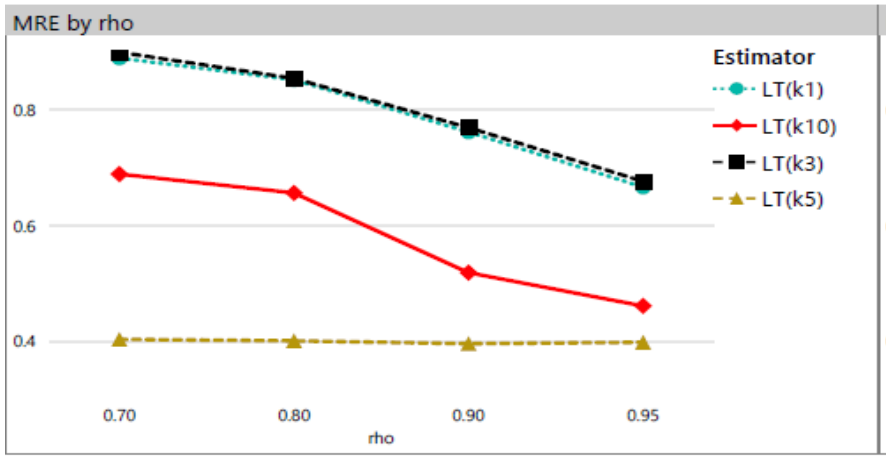


Fig. 1.c: MRE based MSE categorized by degree of correlation when p=6

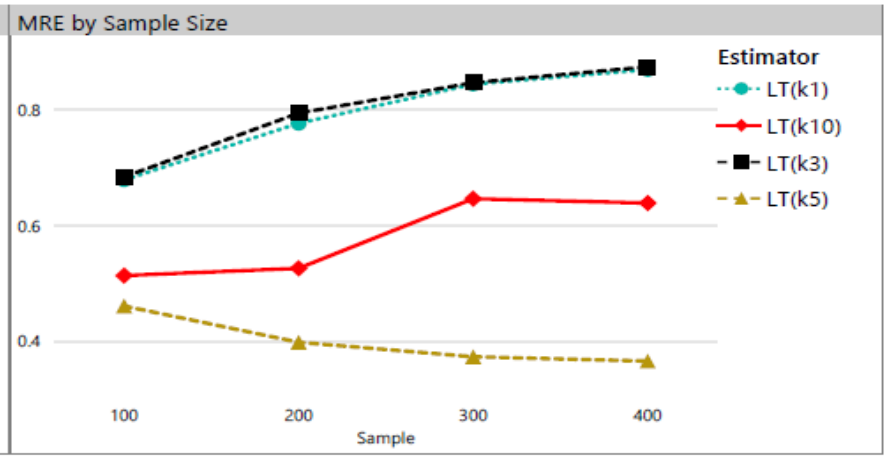


Fig. 1.d: MRE based MSE categorized by sample size when p=6

Fig. 1: MRE based MSE for LTZINB Estimators

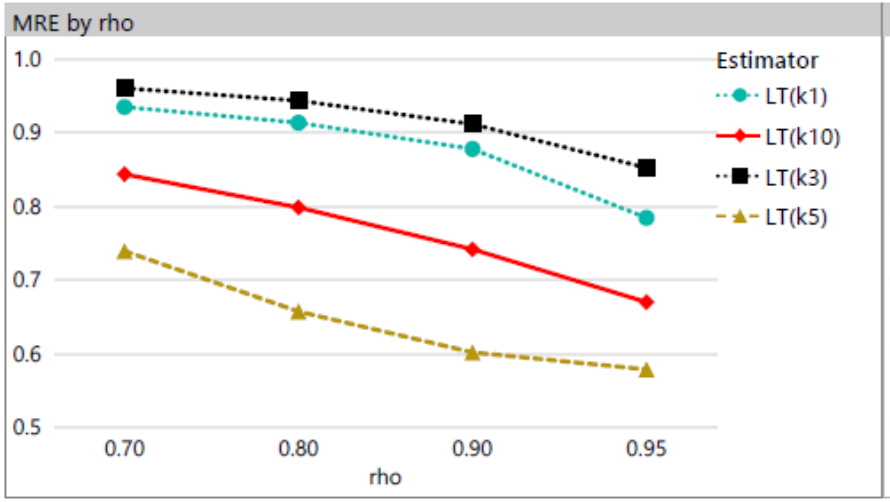


Fig. 2.a: MRE based MAE categorized by degree of correlation when p=2

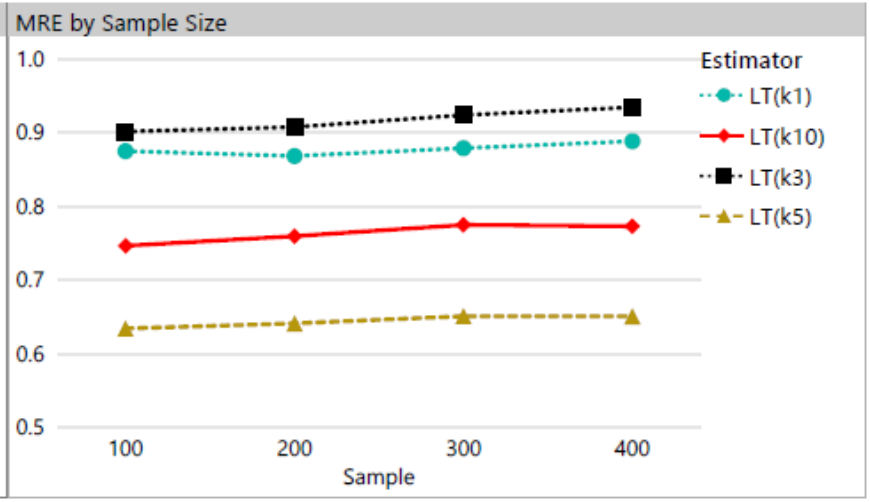


Fig. 2.b: MRE based MAE categorized by sample size when p=2

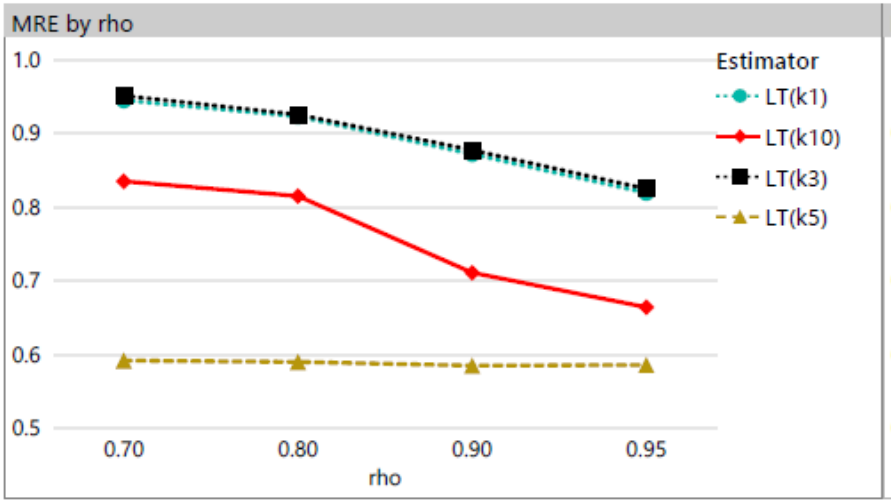


Fig. 2.c: MRE based MAE categorized by degree of correlation when p=6

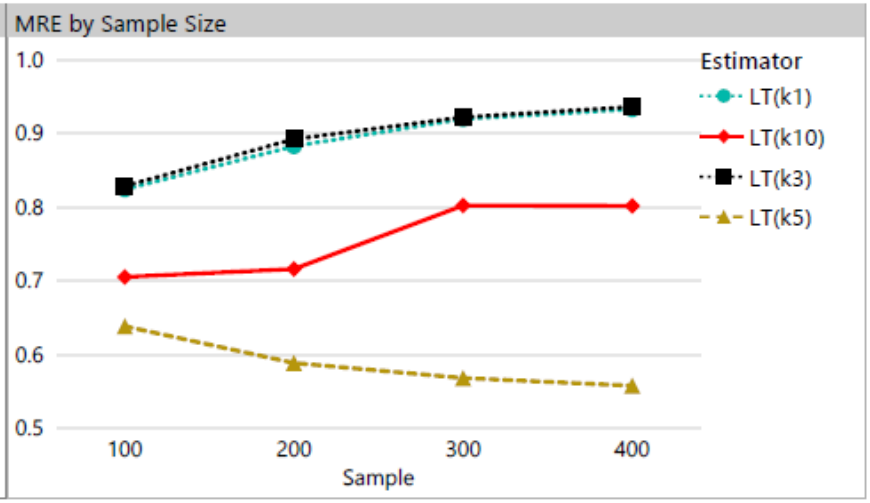


Fig. 2.d: MRE based MAE categorized by sample size when p=6

Fig. 2: MRE based MAE for LTZINB Estimators

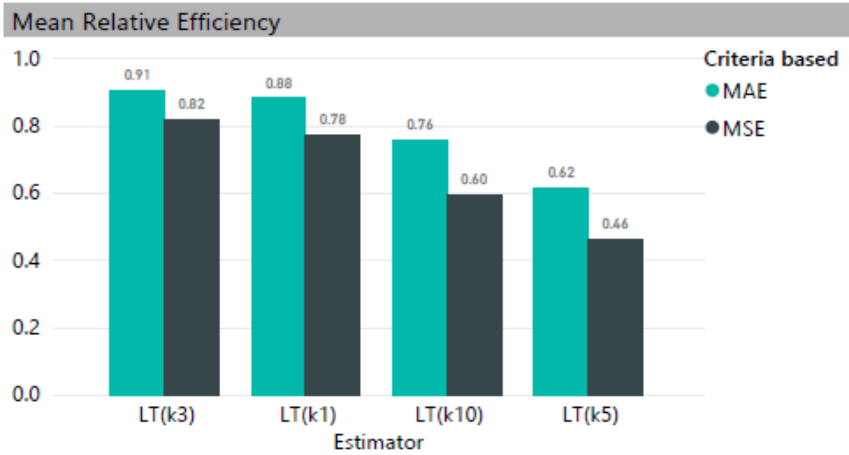


Fig. 3.a: MRE clustered by the used criteria

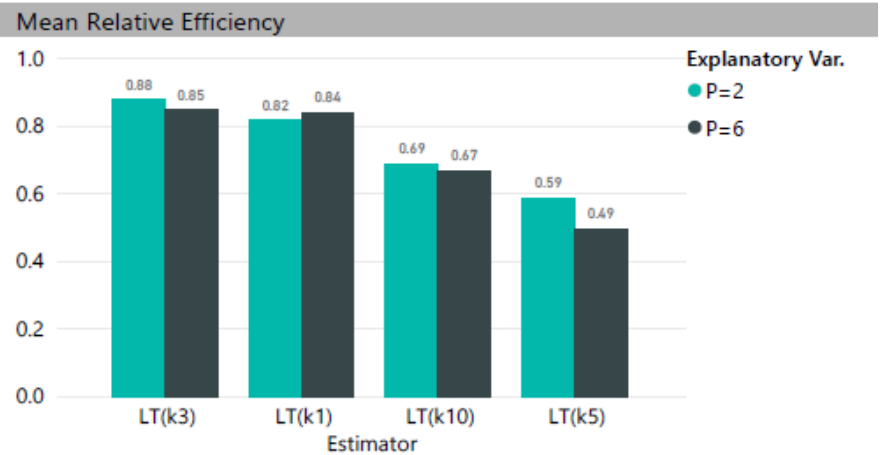


Fig. 3.b: MRE clustered by the number of explanatory variables

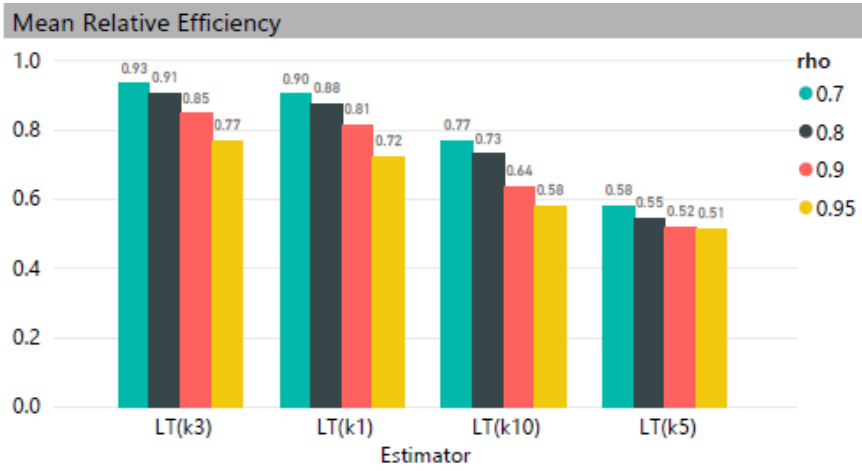


Fig. 3.c: MRE clustered by the degree of correlation

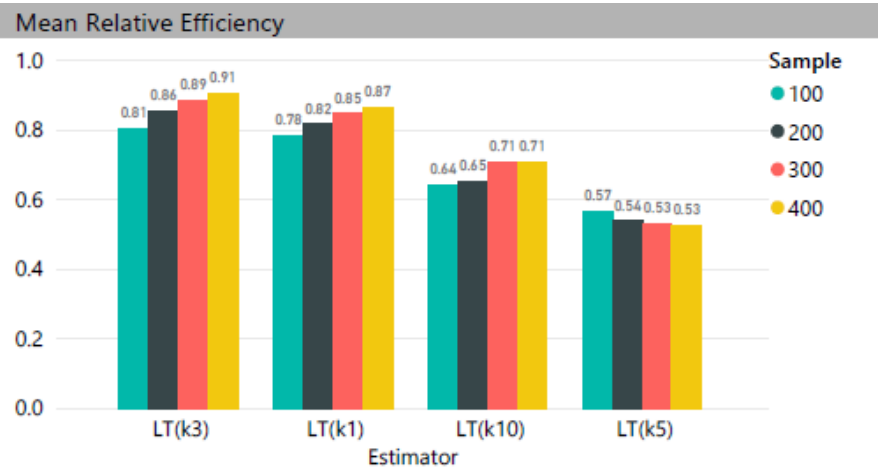


Fig. 3.d: MRE clustered by the sample size

Fig. 3: MRE based MSE and MAE for LTZINB Estimators

5. Conclusion and Future Works

The main problem addressed in this paper is the problem of multicollinearity in ZINB model; some biased estimators have been investigated as a suitable solution to this problem. Below are some conclusions one may see emerging from this work:

- High inter-correlations among the independent variables can seriously affect the ML estimator; since the estimated regression coefficients have high variance and as a result may reduce the precision of estimation.
- The EM algorithm is more preferable when estimating mixture models like ZINB model; since the algorithm is straightforward, the estimates are quite insensitive to the initial values, and the iterations converge monotonically and thus guarantee a good stability of the estimation procedure.
- In a major contribution of this paper, we extend previously proposed shrinkage parameters to the LTZINB modeling, and evaluate the performance of the Liu-type estimator with the new parameters in the simulation study.
- It has been shown in the simulation study that increasing the number of explanatory variables and correlation among them, have negative effect on the performance of the estimators. Whereas, increasing the sample size has positive effect on the performance of the estimators.
- The final conclusion of this paper is that; among biased regression methods, Liu-type estimator has been shown to offer good performance over the other estimators in terms of MSE and MAE criteria especially when using *K5* as a shrinkage parameter.
- The superiority of Liu-type estimator is shown through Monte Carlo simulation study.

Many different adaptations, methods of estimation, and models have been left for the future due to lack of time. Future works concerns other shrinkage estimators of particular

mechanisms, new proposals to try different methods, or simply curiosity. Some ideas can be applied in the future include the following:

- The dispersion parameter maybe taken as a changing factor in the simulation.
- The selected biased estimators can be applied to other count models.
- The new estimators proposed by Asar and Genç (2017), Özkale and Arıcan (2016), Wu and Asar (2016), Asar et al. (2017), Wu et al. (2017), and Wu and Asar (2018) for the logistic model can be extended to count models.
- Similarly, the new estimators proposed by Türkan and Özel (2016), Türkan and Özel (2018) for the Poisson model may be also extended to zero inflated count models.
- Finally, there are plenty of multicollinearity remedies that have not been yet generalized to the area of count models.

References

Abonazel, M. R. (2018). A practical guide for creating Monte Carlo simulation studies using R. *Int. J. Math. Comput. Sci. 4*(1), 18–33.

Abonazel, M. R., & Farghali, R. A. (2019). Liu-Type multinomial logistic estimator. *Sankhya B. (81),* 203-225.

Alkhamisi, M. A., & Shukur, G. (2008). Developing ridge parameters for SUR model. *Communications in Statistics-Theory and Methods, 37*(4), 544–564.

Alkhamisi, M., Khalaf, G., & Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics-Theory and Methods*, *35*(11), 2005-2020.

Asar, Y. (2018). Liu-type negative binomial regression: *a comparison of recent estimators and applications. In: Trends and perspectives in linear statistical inference. Springer, Cham, pp23–39.*

Asar, Y., & Genç, A. (2016). New shrinkage parameters for the Liu-type logistic estimators. *Communications in Statistics-Simulation and Computation*, *45*(3), 1094-1103.

Asar, Y., & Genç, A. (2017). Two-parameter ridge estimator in the binary logistic regression. *Communications in Statistics-Simulation and Computation*, *46*(9), 7088-7099.

Asar, Y., & Genç, A. (2018). A new two-parameter estimator for the Poisson regression model. *Iranian Journal of Science and Technology, Transactions A: Science*, *42*(2), 793-803.

Asar, Y., Karaibrahimoglu, A., and Genc¸ A. (2014). Modified ridge regression parameters: A comparative Monte Carlo study. *Hacettepe Journal of Mathematics and Statistics 43* (5):827–841.

Asar, Y., Arashi, M., & Wu, J. (2017). Restricted ridge estimator in the logistic regression model. *Communications in Statistics-Simulation and Computation*, *46*(8), 6538-6544.

Bhat, S. S. (2016). A comparative study on the performance of new ridge estimators. *Pakistan Journal of Statistics and Operation Research*. *12*(2):317–25.

Dempster, A. P., Schatzoff, M., & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, *72*(357), 77-91.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Papers 94-10, New York University, Leonard N. Stern School of Business, Department of Economics.

Hoerl, A. E., & Kennard, R.W. (1970a). Ridge regression: biased estimation for non-orthogonal Problems. *Technometrics* 12, 55–67.

Hoerl, A. E., & Kennard, R.W. (1970b). Ridge regression: application to non-orthogonal problems. *Technometrics* 12, 69–82.

Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, *4*(2), 105-123.

Huang, J., & Yang, H. (2014). A two-parameter estimator in the negative binomial regression model. *Journal of Statistical Computation and Simulation*, *84*(1), 124-134.

Kaçıranlar, S., & Dawoud, I. (2018). On the performance of the Poisson and the negative binomial ridge predictors. *Communications in Statistics-Simulation and Computation*, *47*(6), 1751-1770.

Kibria, B. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*, *32*(2), 419-435.

Kibria, B. G., Månsson, K., & Shukur, G. (2011). Performance of some logistic ridge regression estimators. *Computational Economics*, *40*(4), 401-414.

Kibria, B. G., Månsson, K., & Shukur, G. (2013). Some ridge regression estimators for the zero-inflated Poisson model. *Journal of Applied Statistics*, *40*(4), 721-735.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14.

Liu, K. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, *22*(2), 393-402.

Liu, K. (2003). Using Liu-type estimator to combat collinearity. *Communications in Statistics-Theory and Methods*, *32*(5), 1009-1020.

Månsson, K. (2012). Developing a Liu estimator for the negative binomial regression model: method and application. *Journal of Statistical Computation and Simulation*, *83*(9), 1773-1780.

Månsson, K., & Shukur, G. (2011). A Poisson ridge regression estimator. *Economic Modelling*, *28*(4), 1475-1481.

Månsson, K., Kibria, B. G., & Shukur, G. (2015). Performance of Some Weighted Liu Estimators for Logit Regression Model: An Application to Swedish Accident Data. *Communications in Statistics-Theory and Methods*, *44*(2), 363-375.

Muniz, G., & Kibria, B. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics—Simulation and Computation*, *38*(3), 621-630.

Nanjundan, G. (2006). An EM algorithmic approach to maximum likelihood estimation in a mixture model. *Vignana Bharathi*, *18*(1), 7-13.

Özkale, M. R., & Arıcan, E. (2016). A new biased estimator in logistic regression model. *Statistics*, *50*(2), 233-253.

Power BI Custom visuals on AppSource. Microsoft AppSource. Microsoft. Retrieved 25 December 2017. URL https://powerbi.microsoft.com/en-us/.

R Development Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org/.

Rady, E. A., Abonazel, M. R., & Taha, I. M. (2018). Ridge estimators for the negative binomial regression model with application. *The 53rd Annual Conference on Statistics, Computer Science, and Operation Research* 3-5 Dec, 2018. ISSR, Cairo University.

Rady, El-Housainy A., Abonazel, Mohamed R., & Taha, I., M. (2019). A New Biased Estimator for Zero-Inflated Count Regression Models. *Journal of Modern Applied Statistical Methods*, Accepted paper (July 2019).

Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, *13*(1), 99-113.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, 197-206, University of California Press, Berkeley, Calif.

Taha, I., M. (2019). Handling multicollinearity problem in generalized linear models. Unpublished master thesis. Faculty of graduate studies for statistical research. Cairo University.

Türkan, S., & Özel, G. (2016). A new modified Jackknifed estimator for the Poisson regression model. *Journal of Applied Statistics*, *43*(10), 1892-1905.

Türkan, S., & Özel, G. (2018). A Jackknifed estimators for the negative binomial regression model. *Communications in Statistics-Simulation and Computation*, *47*(6), 1845-1865.

Wu, J., & Asar, Y. (2016). On almost unbiased ridge logistic estimator for the logistic regression model. *Hacettepe Journal of Mathematics and Statistics*, *45*(3), 989-998.

Wu, J., & Asar, Y. (2018). Performance of the almost unbiased ridge-type principal component estimator in logistic regression model. *Communications in Statistics-Simulation and Computation*, *47*(10), 2925-2937.

Wu, J., Asar, Y., & Arashi, M. (2017). On the restricted almost unbiased Liu estimator in the Logistic regression model. *Communications in Statistics-Theory and Methods*.

Yüzbaşi, B., & Asar, A. (2018). Ridge Type Estimation in the Zero-Inflated Negative Binomial Regression. *Econometrics: Methods & Applications*, page 93.

Zeebari, Z., Shukur, G. & Kibria, B. M. G. (2012). Modified ridge parameters for seemingly unrelated regression model. *Commun. Statist.-Theory Methods 41*(9), 1675–1691.

Appendix A: Shrinkage parameters

Table A.1: Ridge Parameters

S.N	Author	Formulae
1)	Hoerl and Kennard (1970a,b)	$\boldsymbol{K1} = \frac{\hat{\sigma}^2}{\hat{\alpha}^2_{\max}}$
2)	Schaeffer et al (1984)	$\boldsymbol{K2} = \frac{1}{\hat{\alpha}^2_{\max}}$
3)	Kibria (2003)	$\boldsymbol{K3} = \frac{\hat{\sigma}^2}{(\prod_{j=1}^{p} \hat{\alpha}_j^2)^{\frac{1}{p}}}$, $\boldsymbol{K4} = \text{median}\,\{m_j^2\}$
4)	Alkhamisi et al. (2006)	$\boldsymbol{\hat{k}^{KS}_{max}} = \max(s_j)$
5)	Muniz and Kibria (2009)	$\boldsymbol{K5} = \max\left(\frac{1}{m_j}\right)$, $\boldsymbol{K6} = (\prod_{j=1}^{p} \frac{1}{m_j})^{\frac{1}{p}}$, $\boldsymbol{K7} = \text{median}\left(\frac{1}{m_j}\right)$,
6)	Kibria et al (2011)	$\boldsymbol{K8} = \max(m_j)$, $\boldsymbol{K9} = (\prod_{j=1}^{p} m_j)^{\frac{1}{p}}$, $\boldsymbol{K10} = \text{median}(m_j)$, $\boldsymbol{K11} = \max\left(\frac{1}{q_j}\right)$, $\boldsymbol{K12} = \max(q_j)$, $\boldsymbol{K13} = (\prod_{j=1}^{p} \frac{1}{q_j})^{\frac{1}{p}}$, $\boldsymbol{K14} = (\prod_{j=1}^{p} q_j)^{\frac{1}{p}}$, $\boldsymbol{K15} = \text{median}\left(\frac{1}{q_j}\right)$, $\boldsymbol{K16} = \text{median}\,(q_j)$.
7)	Asar et al. (2014)	$\boldsymbol{K17} = \frac{p^2}{\lambda^2_{max}} \frac{\hat{\sigma}^2}{\sum_{j=1}^{p} \hat{\alpha}_j^2}$, $\boldsymbol{K18} = \frac{p^3}{\lambda^3_{max}} \frac{\hat{\sigma}^2}{\sum_{j=1}^{p} \hat{\alpha}_j^2}$, $\boldsymbol{K19} = \frac{p}{(\lambda_{\max})^{\frac{1}{3}}} \frac{\hat{\sigma}^2}{\sum_{j=1}^{p} \hat{\alpha}_j^2}$, $\boldsymbol{K20} = \frac{p}{\left(\sum_{j=1}^{p} \sqrt{\lambda_j}\right)^{\frac{1}{3}}} \frac{\hat{\sigma}^2}{\sum_{j=1}^{p} \hat{\alpha}_j^2}$,
8)	Bhat (2016)	$\boldsymbol{K21}$ = K1 + $\frac{1}{\lambda_{\max} \sum_{j=1}^{p} \hat{\beta}_j^2}$, $\boldsymbol{K22}$ = K1 + $\frac{1}{2\left(\sqrt{\lambda_{\max}/\lambda_{\min}}\right)^2}$
9)	Rady et al. (2019)	$\boldsymbol{K23} = \left[\frac{p\hat{\sigma}^2}{(\lambda_{max}\hat{\beta}_{max})^2}\right]^{1/p}$; $\boldsymbol{K24} = \left[\frac{p\hat{\sigma}^2}{max(\hat{\beta}^2)}\right]^{1/p}$,

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{n-p-1}$, $\hat{\alpha}^2_{\max} =$ is the maximum element of $\hat{\alpha}^2_j$, $m_j = \sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_j^2}}$, $s_j = \frac{t_j \hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_j \hat{\alpha}_j^2}$, t_j is the eigenvalues of the $(X'X)$ matrix, $q_j = \frac{\lambda_{\max}}{[(n-p)\hat{\sigma}^2 + \lambda_{\max}\hat{\alpha}_j^2]}$, $\lambda_{\max}$ = the maximum eigenvalue of $(X'\widehat{W}X)$.

Table A.2: Liu Parameters

S.N	Author	Formulae
1)	Kibria (2003)	$D1 = \max\left(0, \frac{\hat{\alpha}_{\max}^2 - 1}{\frac{1}{\hat{\lambda}_{\max}} + \hat{\alpha}_{\max}^2}\right), D2 = \max\left\{0, \text{median}\left(\frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2}\right)\right\}$
2)	Månsson (2012)	$D3 = \max\left\{0, \frac{1}{p}\sum_{j=1}^{p} \frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2}\right\}, D4 = \max\left\{0, \max\left(\frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2}\right)\right\}$ $D5 = \max\left\{0, \min\left(\frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2}\right)\right\}$
3)	Månsson et al. (2015)	$D6 = \max\{0, \text{median}(q_j)\}, D7 = \max\left\{0, \frac{\sum_{j=1}^{p} q_j}{p}\right\}$ $D8 = \max\{0, \max(q_j)\}, D9 = \max\{0, \min(q_j)\}$

Table A.3: Liu-type Parameters

S.N	Author	Formulae
1)	Asar and Genc (2016)	$d_{Y1} = \text{median}\left(\frac{\lambda_j \alpha_j^2}{2\left(1+\lambda_j \alpha_j^2\right)}\right), d_{Y2} = \frac{\lambda_{\min} \alpha_{\min}^2}{2\left(1+\lambda_{\min} \alpha_{\min}^2\right)}$ $d_{Y3} = \min\left(\frac{\lambda_j \alpha_j^2}{2\left(1+\lambda_j \alpha_j^2\right)}\right), d_{Y4} = \max\left(\frac{\lambda_j \alpha_j^2}{2\left(1+\lambda_j \alpha_j^2\right)}\right)$
2)	Asar and Genc (2018)	$k_{1.\text{AG}} = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{\lambda_j}{\lambda_j \hat{\alpha}_j^2 (1-d)-d}\right), k_{2.\text{AG}} = \text{median}\left(\frac{\lambda_j}{\lambda_j \hat{\alpha}_j^2 (1-d)-d}\right)_{j=1}^{p}$ $k_{3.\text{AG}} = \max\left(\frac{\lambda_j}{\lambda_j \hat{\alpha}_j^2 (1-d)-d}\right)_{j=1}^{p}, k_{4.\text{AG}} = \min\left(\frac{\lambda_j}{\lambda_j \hat{\alpha}_j^2 (1-d)-d}\right)_{j=1}^{p}$ $k_{5.\text{AG}} = \frac{p}{\sum_{j=1}^{p}\left(\frac{\lambda_j \hat{\alpha}_j^2 (1-d)-d}{\lambda_j}\right)}$
3)	Asar (2018)	$k_{\text{AM}} = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{\lambda_j - d\left(1+\lambda_j \hat{\alpha}_j^2\right)}{\lambda_j \hat{\alpha}_j^2}\right), k_{\text{MAX}} = \max\left(\frac{\lambda_j - d\left(1+\lambda_j \hat{\alpha}_j^2\right)}{\lambda_j \hat{\alpha}_j^2}\right)$

Appendix B: Simulation Results

Table B.1: MSE and MAE values of different estimators for **ZINB** model when $\boldsymbol{p = 2}$

	MSE					MAE				
	ML	$LT(K_1)$	$LT(K_3)$	$LT(K_{10})$	$LT(K_5)$	ML	$LT(K_1)$	$LT(K_3)$	$LT(K_{10})$	$LT(K_5)$
$\boldsymbol{\rho = .70}$										
100	0.0495	0.0446	0.0463	0.0356	0.0289	0.2453	0.2367	0.2373	0.2109	0.1789
200	0.0352	0.0276	0.0314	0.0239	0.0203	0.2118	0.1898	0.1991	0.1746	0.1531
300	0.0236	0.02	0.0219	0.0168	0.0143	0.1697	0.1594	0.1636	0.1438	0.1273
400	0.0195	0.0166	0.0184	0.0138	0.0118	0.1579	0.1484	0.1533	0.1332	0.1193
$\boldsymbol{\rho = .80}$										
100	0.1097	0.0889	0.0948	0.0689	0.0591	0.3653	0.3316	0.3386	0.2833	0.2314
200	0.0437	0.0321	0.0376	0.0276	0.0227	0.2338	0.2018	0.2142	0.1842	0.1525
300	0.0308	0.0254	0.0281	0.02	0.0164	0.1967	0.1812	0.1876	0.1576	0.1304
400	0.0207	0.0186	0.0198	0.0141	0.0112	0.1611	0.1551	0.1575	0.1337	0.1097
$\boldsymbol{\rho = .90}$										
100	0.3143	0.214	0.231	0.1717	0.1585	0.6231	0.5185	0.5434	0.4329	0.3696
200	0.0865	0.0686	0.0737	0.05	0.0431	0.3303	0.3014	0.3051	0.2503	0.2002
300	0.0555	0.0407	0.0467	0.0324	0.0274	0.2624	0.2318	0.2415	0.2011	0.159
400	0.0395	0.03	0.0349	0.0225	0.019	0.2267	0.2005	0.211	0.1697	0.1365
$\boldsymbol{\rho = .95}$										
100	0.4924	0.3152	0.3389	0.2559	0.243	0.7897	0.6285	0.6639	0.5183	0.4589
200	0.2418	0.1572	0.1737	0.1287	0.1216	0.5514	0.4424	0.4693	0.3687	0.3219
300	0.1337	0.0818	0.0965	0.0713	0.0663	0.4065	0.3141	0.349	0.2784	0.2378
400	0.0906	0.0546	0.0669	0.0468	0.043	0.3352	0.2572	0.2881	0.2247	0.1894

Table B.2: MSE and MAE values of different estimators for **ZINB** model when $\boldsymbol{p = 4}$

	MSE					**MAE**				
	ML	$LT(K_1)$	$LT(K_3)$	$LT(K_{10})$	$LT(K_5)$	ML	$LT(K_1)$	$LT(K_3)$	$LT(K_{10})$	$LT(K_5)$
$\boldsymbol{\rho = .70}$										
100	0.1984	0.1707	0.1737	0.1286	0.0936	0.7044	0.6559	0.6612	0.5707	0.4446
200	0.086	0.0777	0.0805	0.056	0.0355	0.4644	0.4423	0.4494	0.3779	0.2735
300	0.0527	0.05	0.0508	0.0394	0.0216	0.3646	0.3568	0.359	0.3188	0.2153
400	0.0438	0.0401	0.0423	0.0288	0.018	0.3329	0.3177	0.3262	0.2699	0.195
$\boldsymbol{\rho = .80}$										
100	0.2854	0.2239	0.2364	0.1579	0.1286	0.8461	0.7493	0.7697	0.6219	0.5133
200	0.132	0.112	0.119	0.072	0.0537	0.5729	0.5291	0.5441	0.4224	0.327
300	0.0858	0.0762	0.08	0.0487	0.0338	0.463	0.4382	0.4482	0.351	0.2618
400	0.0571	0.0519	0.0544	0.0352	0.0224	0.3782	0.3618	0.3704	0.299	0.2125
$\boldsymbol{\rho = .90}$										
100	0.6426	0.451	0.4628	0.3108	0.2861	1.2607	1.0687	1.0827	0.86	0.7631
200	0.2156	0.1659	0.1826	0.0974	0.0845	0.7366	0.6522	0.682	0.487	0.4106
300	0.1595	0.127	0.1402	0.0741	0.0631	0.6313	0.5713	0.5972	0.4258	0.3527
400	0.1203	0.1035	0.109	0.0598	0.0453	0.5498	0.5144	0.5267	0.388	0.2977
$\boldsymbol{\rho = .95}$										
100	1.1848	0.6646	0.7188	0.4698	0.4922	1.7276	1.3011	1.36	1.013	1.0116
200	0.4827	0.3225	0.3571	0.1971	0.1946	1.1063	0.9088	0.9569	0.6627	0.6251
300	0.3269	0.2455	0.2591	0.135	0.1211	0.916	0.7973	0.818	0.573	0.4974
400	0.2387	0.1775	0.1973	0.0963	0.0891	0.7774	0.6808	0.7166	0.4759	0.4224

Table B.3: MSE and MAE values of different estimators for **ZINB** model when $\boldsymbol{p = 6}$

	MSE					**MAE**				
	ML	$LT(K_1)$	$LT(K_3)$	$LT(K_{10})$	$LT(K_5)$	ML	$LT(K_1)$	$LT(K_3)$	$LT(K_{10})$	$LT(K_5)$
$\boldsymbol{\rho = .70}$										
100	0.3109	0.2505	0.2548	0.1911	0.1452	1.0789	0.9707	0.979	0.8469	0.6931
200	0.1319	0.1165	0.1192	0.0831	0.0527	0.7088	0.6673	0.6746	0.5659	0.4199
300	0.0833	0.0772	0.0778	0.0623	0.0317	0.5635	0.544	0.546	0.4917	0.3236
400	0.0644	0.0607	0.061	0.0495	0.0236	0.4931	0.4805	0.4819	0.4364	0.2752
$\boldsymbol{\rho = .80}$										
100	0.5251	0.389	0.3907	0.3106	0.2477	1.3944	1.2065	1.2092	1.0742	0.8987
200	0.2068	0.1757	0.177	0.1331	0.0824	0.8822	0.8148	0.8176	0.712	0.5211
300	0.1318	0.1182	0.1185	0.0961	0.0488	0.7028	0.6665	0.6675	0.6033	0.3957
400	0.0952	0.0872	0.0879	0.0642	0.0352	0.6044	0.5776	0.5802	0.4984	0.3395
$\boldsymbol{\rho = .90}$										
100	1.0426	0.65	0.6601	0.4535	0.4813	1.982	1.5672	1.5796	1.2624	1.2688
200	0.4197	0.3137	0.3202	0.1936	0.1687	1.2657	1.0946	1.1059	0.8461	0.7431
300	0.2495	0.2067	0.207	0.1636	0.0907	0.977	0.8909	0.8916	0.7956	0.541
400	0.1942	0.1647	0.1669	0.1026	0.072	0.8582	0.7905	0.7956	0.6221	0.4798
$\boldsymbol{\rho = .95}$										
100	1.8198	1.0037	1.0081	0.7735	0.8113	2.5903	1.9212	1.9255	1.6297	1.6225
200	0.8229	0.5209	0.5408	0.3168	0.3303	1.7549	1.4069	1.4343	1.0361	1.0246
300	0.5275	0.3789	0.3827	0.2364	0.2059	1.4071	1.2013	1.2073	0.9343	0.818
400	0.3869	0.2969	0.2973	0.2286	0.139	1.2121	1.0675	1.0683	0.9372	0.6688

Intra Plot Variation and Multivariate Analysis in Maize Fertilization Experiments.

Abdalla, A.F.M, Okaz, A.M.A, Hager, M.A and Hasham, M.M.

Department of Agronomy, Faculty of Agriculture, Al-Azhar University, Cairo, Egypt.

Abstract

Field experiment in randomized complete block deign with three replications in maize was carried out during the 2017 season. To study the relationship between Kernel yield and associated traits in nitrogen fertilization and varietal experiments in maize using the average per plot (AVP, n=72) and all sample observations (ASO, n=1080). Descriptive statistics, Path coefficient analysis, simple correlation, Multicollinearity diagnosis, Principal component analysis, Multiple linear regression and Stepwise multiple linear regression and were used. Results indicated that the median and mode of the traits Showed a difference trend in some traits. The sub-sample analysis technique is an important way to show the variation within plot segment as a sampling error. Correlation results showed that coefficients among traits in AVP larger than in ASO. The multicollinearity diagnosis by tolerance and variance inflation factor. For two data set the largest correlation between explanatory traits observed between ELA and ELW. In ASO EP traits noncorrelated with KWP. Based on Principal Component Analysis,73.29% in the total variance was accounted for by two principle components included all traits in AVP. In ASO two principle components included all traits except NUL accounted 48.59% of the total variation. indicated that Multiple and Stepwise linear regression indicated that the traits as a predictor of KWP are different form AVP and ASO. Path coefficient analysis has different results for AVP and ASO. These findings could be concluded that using all sample observations helps in studying the variation within the plot and overcome the problem of multicollinearity between the studied traits and to increase confidence in the results obtained.

Keywords: Multivariate Analysis, Plot Variation, Fertilization, Maize.

1. Introduction

Statistics play an important role in experimentation where the experimental design and statistics help to analyze and interpret the results (Box et al. 1978). The importance of statistical science in agriculture is obvious, where the collection, analysis, and interpretation of numerical data are concerned. Statistical principles apply in all areas of experimental work and they have a very important role in agricultural experiments. (Cobanovic,2002).

In agronomic studies, usually the researcher randomly selects a group of plants from each experimental unit to estimate the studied traits or, in other words, in each experimental unit the traits of several plants are assessed, which habitually composes the average of the trait and the use of averages masks individual variances for this experimental unit (Olivoto et al. 2017). The variation among the averages for experimental units will not be affected by sampling error (Vaux et al. 2012).

Maize (Zea mays L.) is the third important major cereal crop after wheat and rice in Egypt (Kandil,2013). In agriculture studied we frequently measure two or more traits on each individual and consequently wiled like to be able to express more precisely the nature of relationship between those traits (Sokal and Rohlf 1995).

Correlation coefficients in general show associations among independent traits and is usually used to measure the relative amount of the influence of each of these traits on the dependent trait, such as grain yield. The proper knowledge of this relationship between grain yield and its traits can significantly improve the efficiency of breeding programs using appropriate selection indicators (Mohamed et al.2008) and (Fard,2014).
When Correlation coefficients between independent variables are high, it can cause problems when you fit the model and interpret the results. This problem is called multicollinearity (Zainodin and Yap 2013). Easily the most common approach to "resolving" situations in which multicollinearity is present is to delete some variables. This deletion is often at the arbitrary discretion of the investigator, where high correlations, low significance, or both, are used as the basis for such a decision (Willis et al. 1978).

Path coefficients to estimate direct and indirect effects of a trait group on a dependent trait. These coefficients are calculated through regression equations using previously standardized variables. However, coefficient estimates can be adversely affected by multicollinearity between traits. Depending on the degree of multicollinearity, the variances associated with path coefficient estimates can assume high but less reliable values (Carvalho et al.2001). In plant sciences the principal component analysis (PCA) has been widely used for the reduction of variables and grouping of genotypes (Kumar et al.2017). Kamara et al (2003) use the PCA to identify traits that accounted for most of variance in maize (Zea mays L.). Stepwise multiple linear regression proved to be more efficient than the full model regression to determine the predictive equation for yield (Naser and Leilah, 1993; Mohamed, 1999).
The present investigation conducted to study the variation within experimental unit (Plant to plant variation) in maize experiments using some statistical method of univariate and multivariate statistical analysis (Descriptive statistics, simple correlation, path coefficient, multicollinearity diagnosis, principal component analysis, multiple linear regression and stepwise multiple linear regression)

2. Materials and Methods

To study the relationship between grain yield and associated traits in nitrogen fertilization and varietal experiments in corn using the average per plot and all samples individual observations. An experiment was carried out during the 2017 summer season.

2.1- Layout of experiment

Factorial experiment (3x8) (3 maize Genotypes × 8 levels of nitrogen) was laid out in Randomized Complete Block Design (RCBD) with three replications, totaling 72 plots. Genotypes were single cross (SC10), three-way cross (TWC321) and open-pollinated cultivar Giza 2. Nitrogen fertilization levels were (0,20,40,60,80,100,120and 140 kg/ fed.). Each plot was included seven rows, 3-m-length, 65 cm spaced. To avoid the prodder effects, the three central rows were assessed. Five plants were randomly sampled and labeled from each plot, totaling 15 plants (observations) for each plot. which were labeled. The traits of plant and ear were measured for each of labeled plant.

2.2- Studied traits and datasets.

The following traits were measured; stem diameter (SD, mm), ear leaf length (ELL, cm), ear leaf width (ELW, cm), ear leaf area (ELA, cm^2), plant height (PH, cm), ear height (EH, cm), ear position (EP, %), number of upper leaves (NUL), ear length (EL, cm), ear diameter (ED, cm), cob diameter (CD, cm), number of rows per ear (NRE), number of kernels per row (NKR), ear weight (EW, g), hundred- kernel weight (HKW,g), kernels weight per plant (KWP,g)(Y). Two datasets were used; first from the average values of the fifteen plants of each plot 72 observation (AVP), the second from values of each plant (plants x replications x Genotypes x Nitrogen levels) 1080 observations (ASO).

2.3-Statistical analysis.

1- Descriptive statistics values of (Mean, median, mode, maximum, minimum, and standard deviation) using (AVP) and (ASO).
2- Correlation matrix of simple correlation coefficients between KWP and other traits were computed as Snedecor and Cochran, (1981). For AVP and ASO cases.
3- Path coefficient analysis was done based on logical relationships between KWP and other traits in the primary and secondary levels of kernel yield to identify direct and indirect path coefficients according to (Dewey et al.1959).
4- Multicollinearity diagnosis among explanatory traits (all studied traits except KWP) for each data set to verify the magnitude of multicollinearity were tolerance and the variance inflation factors (VIF) according to Kutner *et al*. (2005).
5- Principal component analysis was also carried out, as proposed by Jeffers (1967).
6- Multiple linear regression and partial coefficient of determination (R2) was estimated for each yield component (Snedecor and Cochran, 1981) in order to evaluate the relative contribution and to develop the prediction model for KWP (Y).
7- Stepwise multiple linear regression was used according to Draper and Smith (1998) to determine the variable accounting for the majority of total yield variability.

3- Results and Discussion

3.1-Dicriptive statistics

Mean, median, mood, minimum, maximum and Standard deviation were estimated for the studied traits using AVP and ASO are shown in Table 1. The mean had the same value for each trait in AVP and ASO. The median of both EW and KWP was the largest in AVP. The mode was larger for PH, EL, EW and KWP traits in ASO. While, for EH, EP and HKW traits were larger in AVP. The minimum value was smaller while, the maximum and the standard deviation values were larger in ASO for all studied traits. The standard deviation was smaller in the AVP set that is because the use of averages which masks individual variances. The current results agreed with those observed by (Olivoto et al. 2017) who said that the smallest standard deviation observed in the AVP scenario confirmed the first hypothesis that the use of averages masks individual variances.

Table 1. Descriptive statistics (Mean, median, mode, minimum, maximum and Standard deviation) for studied traits estimated using average per plot, n=72 (AVP) and all sample observations, n=1080 (ASO).

Traits‡	*Mean*	*Median*		*Mode*		*Min.*		*Max.*		*S. Dev*	
		AVP	*ASO*	*AVP*	*ASO*	*AVP*	*ASO*	*AVP*	*ASO*	*AVP*	*ASO*
SD	19.91	20.4	20.0	19.4	20.0	13.5	12.0	24.9	29.0	2.5	2.9
ELL	95.35	96.5	96.0	97.3	99.0	75.4	59.0	106.1	115.0	6.4	8.0
ELW	8.87	9.1	9.0	9.20	9.0	6.3	5.0	10.2	12.0	0.8	1.2
ELA	6.37	6.6	6.5	6.40	6.5	3.8	2.7	7.8	9.6	0.9	1.2
PH	255.2	255.2	255.0	223.4	245.0	203.0	130.0	315.3	335.0	25.3	30.2
EH	138.1	135.5	135.0	139.0	135.0	107.7	65.0	177.0	210.0	16.5	20.5
EP	54.03	53.74	53.85	56.2	50.0	48.07	37.5	60.5	73.6	2.4	5.0
NUL	5.97	6.0	6.0	6.7	6.0	4.7	4.0	6.9	9.0	0.5	0.8
EL	17.57	18.1	18.1	17.2	20.0	11.5	7.0	21.7	26.0	2.6	3.3
ED	4.32	4.4	4.3	4.7	4.3	3.5	2.8	5.1	5.7	0.4	0.5
CD	2.25	2.3	2.3	2.2	2.3	1.8	1.4	3.0	3.8	0.3	0.3
NRE	13.25	13.4	14.0	14.4	14.0	10.5	6.0	15.2	22.0	0.9	1.9
NKR	38.89	40.0	40.0	44.9	44.0	21.9	8.0	49.9	57.0	6.6	8.4
EW	190.1	197.7	190.0	197.7	245.0	81.0	45.0	276.0	355.0	46.8	61.4
HKW	35.36	35.00	35.4	34.1	30.1	27.7	17.9	51.8	56.1	4.5	5.8
KWP	157.8	163.4	156.5	155.1	139.0	56.3	30.0	239.2	310.0	42.4	55.6

‡Stem diameter (SD, mm),Ear leaf length (ELL, cm), Ear leaf width (ELW, cm), Ear leaf area (ELA, cm^2), Plant height (PH, cm), Ear height (EH, cm), Ear position (EP, %), Number of upper leaves (NUL), Ear length (EL, cm), Ear diameter (ED, cm), Cob diameter (CD, cm), Number of rows per ear (NRE), Number of kernels per row (NKR), Ear weight (EW, g), Hundred- kernel weight (HKW,g), kernels weight per plant (KWP,g).

3.2- Correlation matrix

The simple correlation coefficients among all studied traits using the AVP (upper diagonal values) and the ASO (down diagonal values) are shown in Table 2. Using AVP set of data it is found 112 out of 120 (93.4%) tested combinations were significant ($p < 0.05$). All traits were correlated with KWP (Y), with the largest correlation coefficient with EW ($r = 0.995$). The largest correlation between explanatory traits observed between ELA and ELW ($r=0.953$).

Table 3 Correlation matrix for the studied traits estimated using average per plot, n=72 (AVP) (upper the diagonal) and all sample observations, n=1080 (ASO) (down the diagonal).

Traits‡	SD	ELL	ELW	ELA	PH	EH	EP	NUL	EL	ED	CD	NRE	NKR	EW	HKW	$KWP_{(Y)}$
SD		.581**	.557**	.600**	.626**	.541**	.067	.634**	.712**	.665**	.455**	.492**	.591**	.712**	.600**	.727**
ELL	.387**		.758**	.918**	.589**	.541**	.162	.554**	.733**	.534**	.360**	.320**	.668**	.710**	.554**	.717**
ELW	.303**	.433**		.953**	.468**	.416**	.091	.478**	.592**	.561**	.513**	.399**	.478**	.606**	.568**	.606**
ELA	.383**	.752**	.916**		.555**	.506**	.141	.550**	.694**	.591**	.484**	.391**	.595**	.696**	.601**	.699**
PH	.471**	.417**	.247**	.359**		.930**	.274*	.624**	.717**	.558**	.380**	.337**	.672**	.702**	.520**	.693**
EH	.383**	.368**	.202**	.309**	.777**		.599**	.611**	.683**	.592**	.426**	.387**	.597**	.673**	.482**	.666**
EP	.013	.064*	.015	.045	-.018	.594**		.282*	.274*	.374**	.310**	.299*	.164	.288*	.160	.287*
NUL	.087**	0.096**	0.063*	0.089**	0.121**	0.107**	0.018**		.667**	.659**	.356**	.539**	.598**	.757**	.503**	.758**
EL	.489**	.439**	.298**	.402**	.493**	.422**	.052	0.142**		.411**	.189**	.359**	.888**	.854**	.677**	.861**
ED	.447**	.321**	.274**	.340**	.377**	.359**	.113**	0.100**	.622**		.768**	.774**	.435**	.767**	.642**	.756**
CD	.315**	.236**	.253**	.288**	.242**	.271**	.133**	-0.012	.318**	.527**		.656**	.051	.468**	.532**	.454**
NRE	.231**	.141**	.170**	.184**	.163**	.177**	.085**	0.070*	.148**	.328**	.255**		.158	.556**	.317**	.537**
NKR	.422**	.418**	.269**	.371**	.478**	.373**	.013	0.100**	.609**	.294**	.065*	.094**		.768**	.522**	.781**
EW	.473**	.431**	.328**	.424**	.491**	.423**	.062*	0.153**	.524**	.467**	.288**	.224**	.487**		.707**	.995**
HKW	.379**	.340**	.271**	.343**	.353**	.324**	.067*	0.076*	.369**	.387**	.355**	.151**	.311**	.607**		.715**
$KWP_{(Y)}$	.479**	.435**	.324**	.423**	.486**	.413**	.053	0.154**	.529**	.462**	.277**	.221**	.489**	.987**	.602**	

‡Stem diameter (SD, mm),Ear leaf length (ELL, cm), Ear leaf width (ELW, cm), Ear leaf area (ELA, cm^2), Plant height (PH, cm), Ear height (EH, cm), Ear position (EP, %), Number of upper leaves (NUL), Ear length (EL, cm), Ear diameter (ED, cm), Cob diameter (CD, cm), Number of rows per ear (NRE), Number of kernels per row(NKR), Ear weight (EW, g), Hundred- kernel weight (HKW,g), kernels weight per plant (KWP,g).

*Significant at the 0.05 level; **Significant at the 0.01 level.

(Y), Kernels weight per plant as a dependent trait.

For the ASO set of data 113 out of 120 (94.1%) tested combinations are values were significant ($p < 0.05$), all traits were correlated with KWP (Y) except EP, with the largest correlation coefficients with EW ($r = 0.987$). The largest correlation between traits were between ELA and ELW ($r=0.916$). These results agreed with these of Alhussein and Idris (2017). whom reported that plant and ear height, ear length and diameter and hundred kernel weight were positively correlated with kernels weight per plant.

Results for correlation indicated that the correlation coefficients estimated with the AVP were larger than those estimated with ASO that is because the average covers up the difference between the individual plants in experimental units for each trait. Thus, the variance values for ASO were largest and correlation coefficients they were smallest for all traits. So, use AVP to estimate correlation coefficient among traits leads to difficulty assessing their relative importance in KWP. Similar results were reported by Blalock (1963) and Hoerl and Kennard (1981).

3.3- Path coefficient analysis

Simple correlation coefficients among kernels weight per plant (KWP) and its explanatory traits were individually partitioned into their component of direct and indirect effects using tow data set (AVP) and (ASO). The results of direct and indirect effect for AVP and ASO are presented in Table 3.

Table 3. Direct and indirect effects for path coefficient analysis for explanatory traits using the average of plot (AVP)and all sample observations (ASO).

Traits	(AVP)			(ASO)		
	Direct	Indirect	Y	Direct	Indirect	Y
SD	0.053	0.674	0.727**	0.014	0.465	0.479**
ELL	0.018	0.698	0.717**	0.010	0.424	0.435**
ELW	-0.012	0.617	0.605**	-0.006	0.330	0.324**
PH	-0.091	0.784	0.693**	0.002	0.484	0.486**
EH	0.057	0.609	0.666**	-0.012	0.426	0.413**
NUL	0.016	0.742	0.758**	0.001	0.153	0.154**
EL	-0.022	0.883	0.861**	0.013	0.516	0.529**
ED	-0.021	0.777	0.756**	-0.001	0.462	0.462**
CD	0.005	0.447	0.453**	-0.010	0.287	0.277**
NRE	-0.014	0.551	0.537**	0.001	0.220	0.221**
NKR	0.047	0.733	0.781**	0.000	0.489	0.489**
EW	0.951	0.043	0.995**	0.976	0.011	0.987**
HKW	0.025	0.691	0.715**	0.005	0.597	0.602**

The results indicated that the most important trait attribution of variation of kernels weight per plant in AVP and ASO where direct effect of EW while the first six traits had a maximum direct effect were EW, EH, SD, NKR, HKW and ELL in AVP and EW, SD, EL, and ELL were the first four traits in direct effect on KWP for ASO. Regarding to indirect effect El had the largest value in AVP (0.883) and HKW had the largest value in ASO (0.597).

In AVP Some traits showed a direct negative effect and low value such as ELW, PH, EL, ED, and NRE although the correlation coefficient between them and the weight of plant grain was positive and significant positive correlation i.e. 0.605,0.693,0.861,0.756 and 0.537, respectively. This result is partial agreement with (Jilo and Tulu 2019) they reported that El exerted a positive indirect effect on grain yield. With respect to ASO ELW, EH, ED and CD showed a direct negative effect while those correlation coefficients were positively significant i.e. 0.324,0.413,0.462 and 0.270. The trait had negative direct effect with high positive indirect effect and positive correlation can be take in account for evaluate the KWP. This result agreed with (Singh and Chaudhary 1977) they stated that in a situation where a trait is having positive association and high indirect effects but negative direct effects, emphasis should be given to the indirect effects and thus, indirect casual factors have to be considered simultaneously for selection. That is cleared for results there is difference between AVP and ASO in achieve results for path coefficient analysis. The current results are in accordance with observed by (Ashmawy,2003).

3.4- Multicollinearity diagnosis.

The results of multicollinearity diagnosis among the explanatory traits for each data set, of (AVP) and (ASO) before and after the remedy were shown in Table 4. It indicated that in a case before remedy; for (AVP) five tolerances closely to zero and seven VIFs more than 10 this for ELL, ELW, ELA, PH, EH, EP and EL traits. For (ASO) two tolerances closely to zero and six VIFs more than 10 this for ELL, ELW, ELA, PH, EH and EP traits. From these results, it is shown high multicollinearity.

The traits ELA and EP had to be removed because the ELA was highly correlated with ELW (r= 0.953**) and (r= 0.916**) for (AVP) and (ASO), respectively. And the trait EP was weakly correlated with KWP (Y) (r= 0.287*) and (r= 0.053) for (AVP) and (ASO), respectively. Alves et al. (2017); Mansfield and Helms (1982) and Montgomery et al. (2012) reported that the multicollinearity may be overcome by excluding the nonadditive traits of the model. This technique depends on the prior diagnosis of the correlation matrix among explanatory traits, adopting procedures that, besides informing the degree of present multicollinearity, also identify the traits that are causing such problems. In a case after remedy multicollinearity problem by excluding the ELA and EP traits.

The variance inflation factor (VIF) was more than 10 for PH (VIF=10.7) in (AVP) with week multicollinearity on the other hand for (ASO) all VIFs less than 5 with no multicollinearity. Form this results it could be summarized that after using variable deletion to remedy the multicollinearity and using all sample observations it could be overcome multicollinearity. Similar results were obtained by Olivoto et al. (2017) and Torres et al. (2015) whom said that multicollinearity can be solved using all sample observations.

Table 4. Multicollinearity diagnosis estimated among original traits and after exclusion ELA and EP using average of plot (AVP) and all sample observation (ASO).

Traits	I				II			
	(AVP)		(ASO)		(AVP)		(ASO)	
	Tolerance	VIF (1)	Tolerance	VIF	Tolerance	VIF	Tolerance	VIF
SD	0.330	3.03	0.595	1.68	0.343	2.91	0.602	1.66
ELL	0.010	105.18	0.024	42.10	0.282	3.54	0.637	1.57
ELW	0.006	177.97	0.009	111.78	0.337	2.97	0.759	1.32
ELA	0.002	484.36	0.005	208.04	Excluded			
PH	0.001	698.37	0.037	26.95	0.093	10.71	0.329	3.04
EH	0.001	1000.74	0.025	40.33	0.104	9.64	0.383	2.61
EP	0.007	144.04	0.062	16.03	Excluded			
NUL	0.337	2.96	0.960	1.04	0.354	2.83	0.961	1.04
EL	0.098	10.24	0.493	2.03	0.101	9.87	0.497	2.01
ED	0.134	7.44	0.543	1.84	0.137	7.30	0.546	1.83
CD	0.209	4.77	0.640	1.56	0.222	4.51	0.641	1.56
NRE	0.266	3.75	0.862	1.16	0.271	3.69	0.864	1.16
NKR	0.111	8.99	0.530	1.89	0.113	8.88	0.536	1.87
EW	0.122	8.21	0.449	2.23	0.128	7.79	0.449	2.23
HKW	0.321	3.11	0.585	1.71	0.327	3.06	0.586	1.71

I, Using all explanatory traits.
II, Using traits after exclusion ELA and EP trait.
(1) Variance inflation factor it invers (tolerance)

3.5- Principal component analysis (PCA).

Principal component analysis was conducted for the explanatory traits for each data set, (AVP) and (ASO). For AVP two principle component were observed in a scree plot for AVP (Figure 1) their eigenvalues were more than 1 with cumulative variability (73.29%) of total variation to kernel yield. For ASO three principle component were observed in a scree plot their eigenvalues were more than 1 with cumulative variability (56.3%) of total variation to kernel yield. Results of principal component analysis are shown in Table 5 it is indicated that, regarding AVP, the first principal component (PC1) explained 60.94% of variation and included all traits except CD and NRE, followed by the second principal component (PC2), which accounted for 12.35.49% of the variation (Table 5). All measured traits were positively loaded onto PC1 and PC2 (Figure 1). The five traits loaded the highest onto PC1 (with scores of >0.30), EW, EL, ED, SD, and PH. With respect to ASO data set the first three principal components had eigenvalues greater than 1. The first principal component (PC1) explained 38.41% of variation and included all traits except CD, NRE and NUL, followed by the second principal component (PC2),

which accounted for 10.18% of the variation and included traits CD and NRE (Table 4). All measured traits were positively loaded onto PC1 and PC2 except NUL trait.

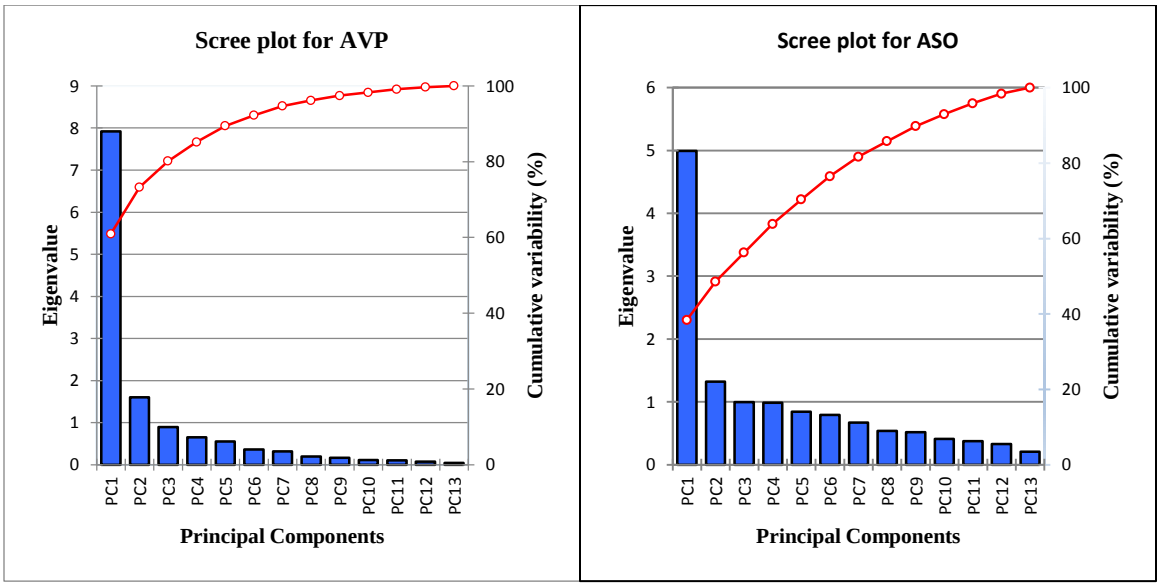


Figure 1. Scree plot for respective eigenvalues and cumulative variability for PCA using AVP and ASO.

Table 5. Principal component analysis for the studied traits, eigenvalues, percent variation and cumulative variance accounted using average of plot (AVP) and all sample observation (ASO).

Components	(AVP)		(ASO)	
	PC1	PC2	PC1	PC2
Eigenvalues	7.92	1.61	4.99	1.32
Variance (%)	60.94	12.35	38.41	10.18
Cumulative (%)	60.94	73.29	38.41	48.59
Component loading				
SD	**0.81**	0.02	**0.70**	0.03
ELL	**0.79**	-0.20	**0.65**	-0.12
ELW	**0.73**	0.07	**0.50**	0.12
PH	**0.81**	-0.24	**0.75**	-0.26
EH	**0.79**	-0.15	**0.68**	-0.19
NUL	**0.79**	-0.02	0.18	-0.19
EL	**0.89**	-0.32	**0.73**	-0.28
ED	**0.84**	0.43	**0.67**	0.42
CD	0.60	**0.69**	0.48	**0.66**
NRE	0.60	**0.63**	0.34	**0.50**
NKR	**0.75**	-0.56	**0.65**	-0.44
EW	**0.93**	-0.07	**0.77**	-0.02
HKW	**0.76**	0.04	**0.64**	0.16

Biplot for traits factor loading for first and second principle component for AVP and ASO dataset are shown in Figure 2. It illustrated that traits had the same trend for AVP and ASO. the first principle component traits were more identical for AVP.

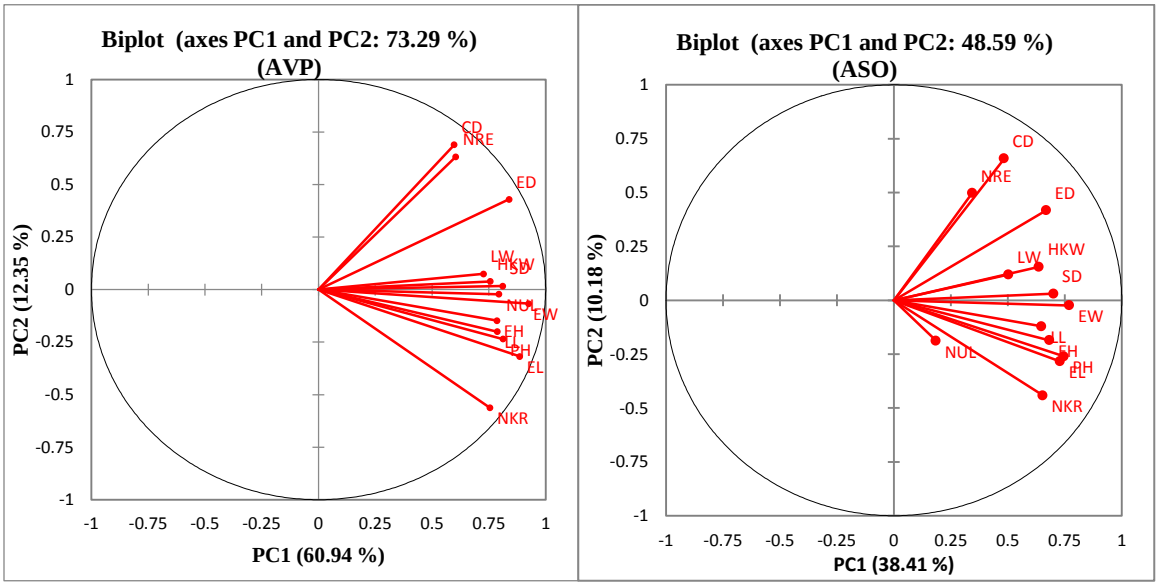


Figure 2. Biplot vectors are trait factor loading for PC1 and PC2 for traits using AVP and ASO.

3.6- Multiple linear regression analysis

Data presented in Table 6 show regression coefficients, standard error and the probability levels of the explanatory traits in predicting KWP for AVP and ASO using full model regression. The obtained results showed that for AVP the prediction equation was formulated as follows:

KWP= -14.94 + 0.92(SD)+0.12(ELL)-0.66(ELW)- 0.15(PH)+0.15(EH)+ 1.26(NUL) -0.36(EL)-2.53(ED) +0.88(CD) -0.64(NRE) +0.31(NKR) + 0.86(EW) +0.23(HKW)

According to this equation, 99.2 % (R^2) of the total variation of KWP could be explained by all explanatory traits. Regrading to the significant traits results showed that SD and PH were significant, and EW was highly significant. This results in disagree with those reported by El-Kadi *et al*. (2011) they stated that the significant traits in model were El, NKR, EP, NRE and HKW. On another hand, in the ASO SD trait was significant, and EW was highly significant. the prediction equation was formulated as follows:

KWP= -18.22 + 0.27(SD)+0.08(ELL)-0.28(ELW)+ 0.01(PH)-0.04 (EH)+ 0.04(NUL) +0.22 (EL)-0.07(ED)-1.71(CD) +0.02(NRE) +0.01(NKR) + 0.89(EW) +0.04(HKW)

With R^2= 97.5 and residual= 2.5. Concerning the goodness of fit of models, results noted that the values of the coefficient of determination (R^2) and adjusted R^2 were equal approximately for each data set the values of root mean square error (RMSE) were 4.127 and 8.183 for AVP and ASO, respectively.

Table 6. Regression coefficients (B), standard error (SE) and significant (Sig.) for explanatory traits of kernels weight per plant using the average of plot (AVP)and all sample observations (ASO).

Traits	(AVP)			(ASO)		
	B	SE	Sig.	B	SE	Sig.
(Constant)	-14.94	15.13	0.33	-18.22	4.25	**0.00**
SD	0.92	0.35	**0.01**	0.27	0.12	**0.03**
ELL	0.12	0.15	0.42	0.08	0.04	0.09
ELW	-0.66	1.12	0.56	-0.28	0.26	0.29
PH	-0.15	0.07	**0.02**	0.01	0.02	0.86
EH	0.15	0.10	0.13	-0.04	0.02	0.11
NUL	1.26	1.61	0.44	0.04	0.14	0.78
EL	-0.36	0.61	0.56	0.22	0.12	0.06
ED	-2.53	3.82	0.51	-0.07	0.77	0.93
CD	0.88	4.09	0.83	-1.71	0.99	0.09
NRE	-0.64	1.03	0.54	0.02	0.15	0.90
NKR	0.31	0.23	0.19	0.01	0.05	0.97
EW	0.86	0.03	**0.00**	0.89	0.01	**0.00**
HKW	0.23	0.20	0.24	0.04	0.06	0.46
Goodness of fit statistics						
	(AVP)			*(ASO)*		
Sig.	0.000			0.000		
R^2	0.992			0.975		
Adjusted R^2	0.990			0.975		

3.7- Stepwise Multiple linear regression analysis

The results of regression analysis by stepwise method for KWP using AVP and ASO are shown in Table 7. For AVP results indicated that EW, SD and NRE traits can justify 99.1 percent of KWP variation. So, this trait is the most important components of KWP in this model. For ASO results indicated that EW, SD, EL and CD traits can justify 97.5 percent of KWP variation. So, this trait is the most important components of KWP in this model. EW as component of corn yield made 98 % in average of the KWP variation for AVP and ASO. Presence of high significant and positive correlation, between EW with KWP shows that the results of stepwise regression are in harmony with correlation results. Furthermore, these results are in line with some earlier findings (Ashmawy, 1998) who stated that EW trait is importance for contribution KWP variation. (Hager *et al*.2012) showed that the number of rows per ear had high relative contribution in KWP according to stepwise regression.

Table 7. Stepwise regression models, regression coefficients(B), standard error (SD) and significance (Sig.) using the average of plot (AVP)and all sample observations (ASO).

Reg. Parameters / Traits	AVP		
	B	SD	Sig.
Intercept	-7.91	7.83	0.32
EW	0.89	0.02	0.00
SD	0.75	0.29	0.01
NRE	-1.37	0.65	0.04
Model sig.	0.00		
R^2	99.10		
Adjusted R^2	99.00		
Prediction equation	KWP = -7.9+0.89(EW)+ 0.75(SD)-1.37(NRE)		
Reg. Parameters / Traits	ASO		
	B	SD	Sig.
Intercept	-15.62	2.39	0.00
EW	0.89	0.01	0.00
EL	0.21	0.10	0.04
SD	0.26	0.12	0.03
CD	-1.85	0.86	0.03
Model sig.	0.00		
R^2	97.50		
Adjusted R^2	97.50		
Prediction equation	KWP = -15.6+0.89(EW)+0.21(EL)+0.26(SD)-1.85(CD)		

4- Conclusion

The current finding could show that the variation within the experimental unit which called Sampling error can affect results of the experiment Specifically if it is significance. Neglecting the differences between observations of the sample (plants from same plot), especially in agronomic experiments, has a great impact on the lack of reliability of the differences between treatments. The use of average data suppresses the individual variation in case of the variation (plant to plant variation) the use of sub sample analysis technique will solve this problem using sampling error as a good measurement for intra plot variation that is for analysis of variance for single variety (Wartha et al. 2016). The use of average will lead to overestimates the magnitude of correlation among traits; thus, the correlation matrices among explanatory traits estimated with average data have the largest multicollinearity. The best strategy to mitigate this problem is to perform the estimates of correlation coefficients with data coming using all samples plants. In addition, the use of multicollinearity diagnosis deletion of some traits will help in application of the suitable multivariate analysis and helps to increase confidence in the results in field crop experiments.

5-References

Alhussein, M. B., & Idris, A. E. (2017). Correlation and path analysis of grain yield components in some maize (Zea mays L.) genotypes. International Journal of Advanced Researcha Publications, 1(1), 79-82.

Alves, B. M., Cargnelutti Filho, A., Burin, C., & Toebe, M. (2017). Linear associations among phenological, morphological, productive, and energetic-nutritional traits in corn. Pesquisa Agropecuária Brasileira, 52(1), 26-35.

Ashmawy, F., & Mohamed, N.A. (1998): Comparision among some statistical precedures in estimating yield and its components in maize.J.Agric.Sci.Mansoura Univ., 23(5):1873-1879.

Blalock, H.M. 1963. Correlated independent variables: The problem of multicollinearity. Soc. Forces 42:233‑237. doi:10.1093/ sf/42.2.233[

Box, G.E.P., Hunter, W.G., & Hunter, J.S. (1978). Statistics for experimenters, an introduction to design, data analysis and model building. New York: John Wiley & Sons.

Carvalho, C. G., Borsato, R., Cruz, C. D., & Viana, J. M. (2001). Path analysis under multicollinearity in S0 x S0 maize hybrids. Crop Breeding and Applied Biotechnology, 1(3).

Cobanovic, K (2002). Role of statistics in the education of agricultural science students. In B. Phillips (Ed.), Proceedings of the Sixth International Conference on Teaching Statistics. Cape Town: International Statistical Institute. CDROM.

Dewey, D. R., & Lu, K. (1959). A Correlation and Path-Coefficient Analysis of Components of Crested Wheatgrass Seed Production 1. Agronomy journal, 51(9), 515-518.

Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). John Wiley & Sons.

El-Kadi, D.A., El-Lakany, M.A.,Abd El-Halim,;/'A.A&Mekhaile , N. E. G. (2010). Statistical models for optimizing productivity of some maize genotypes. Egypt.J.Plant Breed.14(2):115-124(2010)

Fard, F. F., Mirshekari, B., & Amirnia, R. (2014). Multiple regression analysis for studied traits in intercropping of popcorn and cowpea. International Journal of Biosciences (IJB), 4(4), 116-120.

Hager, M. A., El-Hashash, E. F., Zaazaa, E. I., and Fares, W. M. (2012). The interrelationship between grain weight and other ear chracters in maize. Bull.Fac.Agric.Cairo Univ., 63:243-251

Hoerl, A.E., and R.W. Kennard. 1981. Ridge regression−1980: Advances, algorithms, and applications. Am. J. Math. Manage. Sci. 1:5‑83. doi:10.1080/01966324.1981.10737061

Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. Journal of the Royal Statistical Society. Series C (Applied Statistics), 16(3), 225–236.

Jilo,T., & Tulu,L. (2019). Association and path coefficient analysis among grain yield and related traits in Ethiopian maize (Zea mays L.) inbred lines. African Journal of Plant Science, Vol. 13(9), pp. 264-272.

Kamara, A. Y., Kling, J. G., Menkir, A., & Ibikunle, O. (2003). Agronomic performance of maize (Zea mays L.) breeding lines derived from a low nitrogen maize population. The Journal of agricultural science, 141(2), 221-230.

Kandil, E. E. E. (2013). Response of some maize hybrids (Zea mays L.) to different levels of nitrogenous fertilization. Journal of Applied Sciences Research, 9(3), 1902-1908.

Kumar, R., Chikkappa, G. K., Singh, S. B., Mukri, G., Kaul, J., Das, A. K., ... & Bhatia, D. (2017). Multivariate Analysis for Yield and Its Component Traits in Experimental Maize Hybrids. Journal of Agricultural Science, 9(3).

Kutner, M. H., Nachtsheim, C. J., & Neter, J. 2004. Applied linear regression models (4th ed.). New York: McGraw-Hill/Irwin Series.

Mansfield, E.R., and B.P. Helms. 1982. Detecting multicollinearity. Am. Stat. 36:158‑160. doi:10.1080/00031305.1982.10482818.

Mohamed NA (1999) Some statistical procedures for evaluation of the relative contribution for yield components in wheat. Zagazig J Ag Res 26 (2): 281–290.

Mohammed, A. A.; Ali. E. S.; Hamada, A. A.; and Siraj, O. M. (2008). Optimum sowing time for maize Zea Mays) in northern Sudan – Published by Sudan. J. Agric. Res: (2008),12,1- 10.

Montgomery, D.C., E.A. Peck, and G. Vining. 2012. Introduction to linear regression analysis. 5th ed. John Wiley & Sons, Hoboken,NJ.

Naser SM, Leilah AA (1993) Integrated analysis of the relative contribution for some variables in sugar beet using some statistical techniques. Bulletin of the Faculty of Agriculture, University of Cairo 44 (1): 253–266

Olivoto, T., de Souza, V. Q., Nardino, M., Carvalho, I. R., Ferrari, M., de Pelegrin, A. J., ... & Schmidt, D. (2017). Multicollinearity in path analysis: a simple method to reduce its effects. Agronomy Journal, 109(1), 131-142.

Singh, R.K, Chaudhary, B.D. (1977). Biometrical Methods in Quantitative Genetic Analysis. Kalyani Publishers, New Dehli-Ludhiane. 318 p.

Snedecor,G.W. and Cochran, W.G., 1981. Statistical Methods.Seventh.Edition.Iowa State Univ., Press,Ames,USA

Sokal, R., & Rohlf, F. (1995). Biometry. The principles and practice of statistics in biological. New York, USA: Editorial Peter Marshall.

Torres, F. E., Teodoro, P. E., Ribeiro, L. P., Correa, C. C. G., Hernandes, F. B., Fernandes, R. L., ... & Lopes, K. V. (2015). Correlations and path analysis on oil content of castor genotypes. Bioscience Journal, 31(5).

Vaux, D. L., Fidler, F., & Cumming, G. (2012). Replicates and repeats—what is the difference and is it significant? EMBO reports, 13(4), 291-296.

Wartha, C. A., Cargnelutti Filho, A., Lúcio, A. D., Follmann, D. N., Kleinpaul, J. A., & Simões, F. M. (2016). Sample sizes to estimate mean values for tassel traits in maize genotypes. Genet. Mol. Res, 15, 1-13.

Willis, C. E., & Perlack, R. D. (1978). Multicollinearity: effects, symptoms, and remedies. Journal of the Northeastern Agricultural Economics Council, 7(1), 55-61.

Zainodin, H. J., & Yap, S. J. (2013, September). Overcoming multicollinearity in multiple regression using correlation coefficient. In AIP Conference Proceedings (Vol. 1557, No. 1, pp. 416-419). AIP.

An overview of Composite Indices:
A Dimension Reduction Perspective

Asmaa Mohamed Sayed Abdelatif, and Yasmin Mohamed Ibrahim

Faculty of Graduate Studies for Statistical Research, Cairo University

Abstract:

A composite index is a combination of various sources of information known as indicators. These indicators provide a summary of the system that is itself not directly measurable. For the index to be useful and meaningful, its construction requires careful consideration of several important aspects of the potentially disparate and multiple indicators that help convey its meaning, Dobbie and Dail (2013). Composite indices data are high-dimensional, where the number of variables may exceed the number of observations. A composite index is a single number summary for each observation in the data. No matter how informative a single number is, it cannot capture all the features of such highly dimensional data. One purpose of this paper is to look at these data from various angles using different techniques in an attempt to discover and extract the wealth of information or knowledge that these data contain. We particularly focus our attention on three challenging issues: (a) The choice of weights for indicators, (b) How to avoid measuring the standard error of the index, and (c) High dimensionality of the data. We look at the data from a multivariate point of view and use existing techniques (such as the Principal Components Analysis and Multidimensional Scaling). We illustrate these methods using the Economic and Social Rights Fulfillment Index (ESRFI).

Key Words: Bootstrap; Composite Indices; Data Quality; Dimension Reduction; Indicator Weights; Margin of Error; Missing Observations; Multidimensional Scaling; Outliers; Principal Components Analysis

1. Introduction

The construction of a composite index usually starts with defining the concept of interest that needs to be measured. For example, the Corruption Perception Index (CPI), the Gender Inequality Index (GII), the Ibrahim Index of African Governance (IIAG), the Economic and Social Rights Fulfillment Index (ESRFI), the Arab Knowledge Index (AKI), and the International Knowledge Index (IKI), to mention just a few.

These concepts are difficult to measure directly. So, experts in the domain field select several variables to be used as proxy measures of the concept of interest. Some of these variables are grouped into indicators.

An indicator is usually a weighted sum (e.g., the mean) of the variables included in it. Thus, an indicator can be thought of as an index of the variables involved. For simplicity the terms variables and indicators will be used synonymously. When constructing composite indices, several challenges appears. The following are examples of these challenges and the most common method of handling it:

1.1 Data Availability and Quality

It is important to use good and well-known data providers in the construction of the composite indices. The data sources can provide raw or modified data. The variables should be selected based on their relevance, accessibility and timeliness. Proxy measures can be used if the desired data were not available. The accuracy of proxy measures can be investigated through sensitivity analysis. Sometimes people tend to use what is available, but poor data quality will produce poor analyses. A good composite index should allow the improvement and development of its method of construction and renew its data (Mazziotta, 2013).

The Economic and Social Rights Fulfillment Index (ESRFI) is constructed and computed for the first time in Egypt by Ahmed (2013). As the name suggests, the index is designed to measure the extent of the fulfillment of the economic and social rights. The index is based on the 2010 Egyptian Household Conditions Observatory Survey (EHCOS). The index is first computed on the individual level then aggregated by 24 Governorates in Egypt. The survey was distributed
at all governorates except for frontier governorates. The survey has 44,039 individuals, which represent 10550 households. The ESRFI data consists of 76 variables, which are grouped into the 35 indicators as shown in Table 1.

We look at the data from a multivariate point of view and use existing techniques (such as the Principal Components Analysis and Multidimensional Scaling). These methods will be illustrated using the Economic and Social Rights Fulfillment Index (ESRFI).

Table 1: The Structure of ESRFI

Dimension	Indicator Number	No. of variables
Right to food	1	14
	2	1
	3	1
	4	1
	5	14
Right to Education	6	1
	7	1
	8	3
	9	1
Right to Adequate Housing	10	1
	11	1
	12	1
	13	1
	14	2
	15	8
	16	1
Right to Health	17	2
	18	1
	19	1
	20	1
	21	1
	22	1
Right to Decent Work	23	1
	24	1
	25	1
	26	1
	27	1
	28	1
	29	5
	30	1
	31	1
	32	1
	33	1
	34	1
	35	1

Total		76

1.2 Different Indicator Scales

Because the data comes from different sources and in different units of measurements, normalization is required before any aggregation. The following are normalization methods mentioned in OECD (2008):

1. Ranking, which is the simplest method of normalization, is not affected by outliers. This method allows the performance of countries to be compared over time based on its position or rank.
2. Standardization (*Z*-score); This method converts the indicators to a common scale with 0 mean and standard deviation 1 (De Vaus, 2002).
3. Min-Max normalization causes all indicators to have equal range from 0 to 100. This can be done by subtracting the minimum of the indicator from each observation then dividing by the range of the indicator. The main disadvantage of this method is that it is affected by the presence of outliers.
4. Distance to a reference method measures the relative position of an indicator to a reference point. The reference point could be a specific number, an external benchmarks country or the average country of the group. The reference point receives a score of 1, while the others receive scores depending on their position to the reference point. This method is also affected by outliers.
5. Categorical scale method assigns a score to each indicator. This score depends on the percentiles of the distribution of the variable or an indicator across countries. For example, the top 5 percent we will get a score of 100 and the worst 5 percent we will get a score of 0. The main advantage of using this method is that it rewards the best performing countries and penalizes the worst performing countries. A disadvantage of using this method is that it ignores a large amount of information about the variance of the transformed indicator.
6. The relevance to the mean method gives countries around the mean 0, while countries below the mean receive –1, and countries above the mean receive 1. This method is simple and not affected by the outliers, but it is criticized because for example if the country X is 25 percent below the mean and country Y is 70 percent below the mean they will both receive the same score.
7. The managerial evaluation method allows every manger in the firm gives a score to the performance of his company.
8. Percentage of annual differences over years. This method considers the percentage growth with respect to previous years. This kind of transformation can be used only if the data were available over a number of successive years.

The Variables in the ESRFI data were normalized using the min-max method as follows:

$$X \leftarrow \frac{X-\min(X)}{\max(X)-\min(X)} * 100 \qquad (1)$$

1.3 The Presence of Missing Values in the Data

The presence of missing values threatens the validity of the statistical inference. There are various methods of handling missing data like complete or available case analysis and the overall mean imputation but the main disadvantage about these methods that they lead to inefficient analysis (Donders et al. (2006)). A more reasonable method is multiple imputation. The selection of the imputation method is related to the type of the missing observations. Missing observations
an be categorized as follows (Roderick and Donald 1987):

1. Missing at random. Missing values don't depend on the variable itself, but they are conditional on other variables in the data set.
2. Missing completely at random. Missing values are not related to the variable or any other variables in the data. Like losing a questionnaire of a study accidentally. Statistical inference in that case will not be biased but of course it will be less efficient.
3. Not missing at random. Missing values depends on the values themselves. Like people of high income tends not to report their income.

The methods of handling missing values are:

1. Case deletion or complete case analysis: Delete the missing record and do the analysis based on the rest of the data. In this case we end up with less information which causes the standard error to increase in the reduced sample (Burren, 2012).
2. Multiple imputations: It provide several values for each missing value. It can more effectively represent the uncertainty due to the imputation. Most of the statistical packages nowadays offer the multiple imputation technique like Monte Carlo simulation (Stephen, 1998 and Allison, 2000).
3. Single imputation: Direct replacement of missing values by new values from the target population. Like mean/median/mode substitution, regression imputation, and EM (expectation-maximization)

imputation. Single imputation underestimates the variance, because it affects the uncertainty of the imputation (Schimitt et al., 2015 and Zhang, 2016). For the ESRFI data missing values were replaced by the median.

1.4 The Presence of Outliers

An outlier in a variable is defined to be an observation that lies an abnormal distance from other values of the variable. Methods for the detection and treatment of outliers are different from one index to another. Recently some of these indices are following the multivariate methods to detect outliers (Rocke and Woodruff, 1996). These methods include:

1. Mahalanobis Distance (see, e.g., Hadi (1992, 1994) and Kannon, 2015).
2. BACON Algorithm (Rocke and Woodruff (1996) and Billor, Hadi, and Velleman (2000)).

If outliers exist in the data then Robust Estimation could be used (Atkinson, 1994 and Penny, and Jolliffe, 2002).

Ahmed (2013) examined the ESRFI data and found them to be free from multivariate outliers.

1.5 The Choice of Indicator Weights

Weights can have a significant effect on the overall result of a country performance or ranking. A number of weighting technique exists such as (Dobbie and Dail, 2013):

1. Equal Weights. This approach is commonly used based on the hypothesis that all indicators are equally important (Booysen, 2002).

2. Data driven weighting schemes. This approach adapts the idea of setting weights based on the dimensions of the data itself such as (Decanco and Lugo, 2013):

(a) Frequency based weights. Weights based on the proportional of population that suffering deprivation in that dimension. For example, the smaller proportion of individuals that suffer from a certain deprivation will take a higher weight.

(b) Most favorable weights. A researcher might select the most favorable weighting scheme for each variable (Zhou, Ang, and Poh, 2007).

(c) Regression based weight. Weights are assigned based on data using regression methods (see, Bernini, Guizzardi, and Angelini, 2013).

3. Normative weights surveys directly how the individuals would tradeoff different indicators of wellbeing. Or to use budget allocation so the number of representative groups are asked to distribute a budget of points to a number of dimensions paying more for those dimensions whose importance they want to be highlighted.

The weights for SSRFI, computed by Ahmed (2013) were based on an opinion survey of 1002 individuals. These weights are shown in the Table 2.

Table 2: Weights Used for the Computation of ESRFI

Dimension	Food	Education	Adequate Housing	Health	Decent Work
Score	3.85	2.53	4.08	2.41	3.71
Weight	0.232	0.152	0.246	0.145	0.224

1.6 Margin of Error

Composite index numbers are subject to various types of errors such as the use of proxies due to the difficulties in obtaining direct measurements; the variability in the methodology and completeness of data sources; etc. Therefore, index numbers should be accompanied by error or confidence bands that reflect their imprecision. For example, the economic and social rights fulfilment (ESRFI) comes with error bands for each of the Governorates' score. These bands

are calculated using an approach called Bootstrap (Efron and Tibshirani, 1993 and Davison and Hinnkley, 2008). Bootstrap error bands are nonparametric, that is, no assumptions about the probability distribution of the data are made. However, bootstrap has its own disadvantages:

a. It is computationally intensive because it requires a huge number of bootstrap samples for the resulting error bands to be reliable.
b. The resulting error bands are random, that is, it gives different error bands every time the method is applied to the same data.
c. The bootstrap assumes that the indicators are interchangeable within their respective subcategories. These disadvantages cast doubt as to the appropriateness and reliability of using bootstrap. The bootstrap ESRFI results are shown in Table 3.

2. The Use of Dimensional Reduction Techniques

Statistical and machine reasoning methods face a rough problem when dealing with high-dimensional data, and normally the number of input variables is reduced before a data analysis algorithm can be successfully applied. The dimensionality reduction can be made in two different ways: by only keeping the most relevant variables from the original dataset (this technique is called feature selection) or by

exploiting the redundancy of the input data and by finding a smaller set of new variables, each being a combination of the input variables, containing the same information as the input variables (this technique is called dimensionality reduction) (Sorzano, et al.2014).

Table 3: Economic and Social Rights Fulfillment Index for Egypt by 24 Governorates

				95% Confidence	
Governorates	Mean	Rank	Std.Error	Lower	Upper
Giza	68.11	1	0.22	67.69	68.53
Alexandria	68.03	2	0.19	67.64	68.43
Cairo	67.86	3	0.17	67.52	68.18
Port Said	67.02	4	0.31	66.39	67.63
Dametta	66.52	5	0.26	65.98	67.03
Helwan	65.96	6	0.24	65.46	66.44
Suez	65.05	7	0.27	64.56	65.58
Al Gharbia	64.76	8	0.19	64.36	65.12
Al Ismailia	64.05	9	0.27	63.55	64.58
Al Dakahila	63.38	10	0.18	63.04	63.73
Quena	63.2	11	0.22	62.75	63.65
Al Kaliubia	62.41	12	0.19	62.04	62.77
Aswan	62.73	13	0.25	61.91	62.88
Al Menofia	62.31	14	0.18	61.94	62.65
Al Sharkia	61.46	15	0.17	61.12	61.8
Six of October	61.21	16	0.22	6.76	61.62
Al Behera	60.74	17	0.18	60.39	61.09
luxor	60.56	18	0.27	60.01	61.07
Menia	60.23	19	0.19	59.88	60.6
Bani Suef	59.42	20	0.2	59.03	59.83
Al Fayoum	59.09	21	0.18	58.73	59.46
Kafr Al Sheikh	58.89	22	0.21	58.47	59.32
Sohag	58.55	23	0.19	58.19	58.9
Assuit	57.71	24	0.17	57.37	58.06
Total	62.69			62.6	62.77

Ahmed (2013)

2.1 Multidimensional Scaling Analysis

Multidimensional Scaling (MDS) is a dimension reduction technique but works on reverse order. Suppose we have a multivariate data set **X** consisting of n observations on p variables. From **X** we compute an $n \times n$ matrix $\mathbf{D_X}$ whose elements are the Euclidean distances (dissimilarities) between each pair of the observations (rows) in **X**. We wish to find an $n \times k$ matrix **M** such that the distances $\mathbf{D_M}$,

obtained from **M**, are approximately equal to the Euclidean distances in $\mathbf{D_X}$. In this case, **M** is said to be geometrically similar to **X**. This technique is known as Multidimensional Scaling and the variables in **M** are called the Principal Coordinates (Kruskal and Wish, 1978).

It can be shown (see, e.g., Manly, 2004) that the variables in **M** are obtained by

$$M = V_k \Lambda_k^{-1/2}, \qquad (2)$$

where **V** is the first *k* eigenvectors corresponding to the largest *k* eigenvalues of the $n \times n$ matrix XX^T and $\Lambda_k^{-1/2}$ is the square root of the diagonal matrix with its diagonal elements are the largest *k* eigenvalues of XX^T. The columns of **M** are called the Principal Coordinates. The parameter *k*, which is the number of variables in **M**, is chosen by the user. We illustrate using two different choices for *k*. First, the adequacy of the *k*-dimensional representation of **M** can be judged by the proportion

$$Pk = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \qquad (3)$$

Values of P_k of about 0.95 suggest a reasonable MDS representation. Accordingly, *k* is the smallest integer such that $P_k \geq 0.8$. Second, *k* could be the smallest integer such that $\lambda_k \geq \varepsilon$, where ε is a small positive value (e.g., 0.00001). We use this latter value for *k*.

Accordingly, the computations of the Multidimensional Scaling can be summarized as follows:

1. Compute the ordered eigenvalues $(\lambda_1, \dots, \lambda_n)$ and the corresponding eigen-vectors $(= V_1, \dots, V_n)$ of the matrix XX^T.
2. Choose *k* using one of the above methods.
3. Compute **M** as in (2)
4. If the correlation between any column of **M** and ESRFI is negative, then replace that column by its negative. One can create two indices based on **M**.

2.1.1 The Multidimensional Scaling Index (MDSI)

An index could be derived from the Principal Coordinates in **M** in (1) as follows: We first compute the mean of the rows of the Principal Coordinates in **M**, denoted by $\overline{M}$. Then transform $\overline{M}$ into an index-like. using the following formula:

$$MDSI_i = \min(ESRFI) + \frac{\overline{M_i} - \min(\overline{M})}{Range(\overline{M})} * Range\ (ESRFI), \quad i = 1, \dots, n, \qquad (4)$$

where $MDSI_i$ is the index of the *i*th Governance, so that the MDSI will have the same range as that of ESRFI. The MDSI then satisfies $\min(ESRFI_i) \leq MDSI_i \leq \max(ESRFI_i)$.

2.1.2 The Weighted MDS Index (WMDSI)

The MDSI in (4) is based on the unweighted Principal Coordinates in **M**. The columns in **M**, however, do not have equal variances. They can then be weighted by their variances, which are $\lambda_1, \lambda_2, \ldots, \lambda_k$, Therefore, we replace **M** in (2) by $W = M\Lambda$ where **Λ** is a diagonal matrix with λ_j on its *j*th diagonal element. If any column in **W** has a negative correlation with the ESRFI, we multiply it by –1. We then compute the mean of the rows of the Weighted Principal Coordinates in **W**, denoted by $\bar{W}$. Finally, we transform the vector $\bar{W}$ and obtain the Weighted Principal Components Index (WMDSI) as follows:

$$WMDSI_i = \min(ESRFI) + \frac{\bar{W}_i - \min(\bar{W})}{Range(\bar{W})} * Range\ (ESRFI), \qquad i = 1, \ldots, n, \qquad (5)$$

where $WMDSI_i$ is the index of the *i*th country, so that the WMDSI will have the same range as that of the ESRFI. Accordingly, the WMDSI index-like scores satisfying

$$\min(ESRFI_i) \leq WMDSI_i \leq \max(ESRFI_i).$$

2.2 Principal Components Analysis (PCA)

The Principal Components Analysis (PCA) is a very well know technique for dimension reduction (see, e.g., Johnson and Wichern (1992) and Srivastava, (2002), Rencher and Christensen (2012)). Given a data matrix **X** of dimension *n* observations on *p* usually highly correlated variables (as is the case with composite indices data), we wish to find a set of uncorrelated (orthogonal) variables **Y** such that the variables in **Y** are linear functions of the variables in **X**. necessarily nonnegative. Notwithstanding this note, we create three new indices based on **Y**. This is explained in the following sections.

2.2.1 The Principal Components Index (PCI)

An index could be derived from the Principal Components in **Y** in (6) as follows: We note first that the normalized eigenvectors are unique up to sign. That is, if V_j is a normalized eigenvector, then $-V_j$ is also a normalized eigenvector. We therefore choose the sign of V_j to make the correlation between Y_i and another index such as the ESRFI positive. Second, we compute the mean of the rows of the Principal Components in **Y** (after reversing the signs when necessary), which we denote by $\bar{Y}$. We also

note here that the vectors in $\overline{Y}$ do not necessarily lie between 0 and 100 and hence, the vector $\overline{Y}$ is not an index-like. It can, however, be transformed into any desired range using the following formula

$$PCI_i = \min(ESRFI) + \frac{\bar{Y}_i - \min(\bar{Y})}{range(\bar{Y})} * range\ (ESRFI) \quad i = 1, \dots, n, \qquad (7)$$

where PCI_i is the index of the *i*th country, so that the PCI will have the same range as that of ESRFI . Of course, any other index (such as the EWI) can be used in the above formula instead of the ESRFI. Consequently, this makes PCI index-like scores satisfying min ($ESRFI_i$) ≤ PCI_i ≤ max($ESRFI_i$).

2.2.2 The Weighted Principal Components Index (WPCI)

The PCI in (7) is based on the unweighted Principal Components in **Y**. The columns in **Y**, however, do not have equal variances. They can then be weighted by their variances, which are the ordered eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$, Therefore, we replace **Y** in (6) by

$$\mathbf{Z} = \mathbf{Y}\mathbf{\Lambda}; \qquad (8)$$

where **Λ** is a diagonal matrix with λ_j on its *j*th diagonal element. If any column in **Z** has a negative correlation with the ESRFI, we multiply it by –1. We then compute the mean of the rows of the Weighted Principal Components in **Z**, denoted by $\bar{Z}$. Finally, we transform the vector $\overline{Z}$ and obtain the Weighted Principal Components Index (WPCI) as follows:

$$WPCI_i = \min(ESRFI) + \frac{\bar{Z}_i - \min(\bar{Z})}{range(\bar{Z})} * range\ (ESRFI), \quad i = 1, \dots, n, \qquad (9)$$

where $WPCI_i$ is the index of the *i*th country, so that the WPCI will have the same range as that of the ESRFI. Accordingly, the WPCI index-like scores satisfying min ($ESRFI_i$) ≤ $WPCI_i$ ≤ max(($ESRFI_i$).

Summary and Conclusions

The contributions of this paper can be summarized as follows: As an alternative methodology to computing composite indices, we propose the use of multivariate technique especially the dimension reduction techniques to rank objects or Governates. The use of dimensional reduction techniques avoids two issues with composite indices, that is, the need to group the variables or indicators by experts and the need to compute the standard errors. Governates in the same group are taken to be similar (having approximately the same rank) and Governorates that are classified into different groups are interpreted as having different ranks.

The second-best alternative is the WMDI and WPCI because they are the ones with the large standard deviation. These conclusions, however, is limited to this data set, which calls for more in-depth comparisons, for example, using large scale simulation studies

Table 4: The ESRFI and the Four Proposed Indices for ESRFI Data.

No.	Governorates	ESRFI	PCI	WPCI	MDSI	WMDSI
1	cairo	78.8	78.7	78.7	78.7	78.7
2	Alexandria	73.9	72.9	75	72.9	75
3	Port Said	73	70.6	73	70.6	73
4	Suez	68.1	68.8	69.4	68.8	69.4
5	Helwan	65	64.3	66.3	64.3	66.3
6	Six of October	48.4	53.7	51.9	53.7	51.9
7	Dametta	66.2	65.3	68.5	65.3	68.5
8	Al Dakahila	56.2	58.8	57.2	58.8	57.2
9	Al Sharkia	50.7	53.5	51.2	53.5	51.2
10	Al Kaliubia	56.2	55	63	55	63
11	Kafr Al Sheikh	41.5	37.2	42	37.2	42
12	Al Garbia	65.7	58.5	66	58.5	66
13	Al Menofia	57.4	46.6	56.7	46.6	56.7
14	Al Behera	46.9	52	49.2	52	49.2
15	Al Ismailia	59.5	46.9	62.9	46.9	62.9
16	Giza	73.7	75.8	77.6	75.8	77.6
17	Bani Suef	44.8	66	41.8	66	41.8
18	Al Fayoum	39.8	52.4	38.8	52.4	38.8
19	Menia	45.6	50	44.1	50	44.1
20	Assuit	40	55.7	37.2	55.7	37.2
21	Sohag	38.8	51.4	40.3	51.4	40.3
22	Quena	64.8	63.4	61	63.4	61
23	Aswan	54.2	58.2	53.5	58.2	53.5

24	Luxor	44.9	42.7	48.7	42.7	48.7

Table 5: Summary statistics of ESRFI and the Four Proposed Indices for ESRFI Data

Variable	Mean	Median	StdDev	Min	Max
ESRFI	56.4	56.3	12.4	38.8	78.8
PCI	58.7	57	10.6	37.2	78.7
WPCI	57.3	57	12.9	37.2	78.7
MDSI	58.3	57	10.6	37.2	78.7
WMDSI	57.3	57	12.9	37.2	78.7

Table 6: The Correlation matrix of ESRFI and the Four Proposed Indices for ISRFI

Index	ESRFI	PCI	WPCI	MDSI	WMDSI
ESRFI	1	0.778	0.983	0.778	0.983
PCI	0.778	1	0.725	1	0.725
WPCI	0.983	0.725	1	0.725	1
MDSI	0.778	1	0.725	1	0.725
WMDSI	0.983	0.725	1	0.725	1

References:

Ahmed, E. R. M. (2013), "Constructing a Rigorous Economic and Social Rights Fulfillment Index for Egypt," M.Sc. Thesis, Faculty of Economics and Political Science, Cairo University.

Allison, P. D (2000), "Multiple Imputation for Missing Data: A Cautionary Tale," *Sociological Methods and Research*, 28, 301–309.

Atkinson A. C. (1994), "Fast Very Robust Methods for The Detection of Multiple Outliers," *Journal of The American Statistical association*, 89, 1329–1339.

Bernini, C., Guizzardi, A., and Angelini, G. (2013), "DEA-like Model and Common Weights Approach for the Construction of a Subjective Community Well-being Indicator," *Social indicators research*, 114(2), 405–424.

Billor, N., Hadi, A. S., and Velleman, P. F. (2000), "Blocked Adaptive, Computationally-Efficient outlier Numerator," *Computational Statistics and Data Analysis*, 34, 279–298

Booysen, F. (2002), "An Overview and Evaluation of Composite Indices of Development," *Social Indicators Research*, 59, 115–151.

Burren, S. V. (2012), *Flexible Imputation of Missing Data*, New York: Chapman and Hall.

Davison, A. C., and Hinkley, D. V. (2008), *Boot Strap Methods and Their Application*, New York: Cambridge University Press.

Decanco, K., and Lugo, A. M. (2013), "Weights in Multidimensional Indices of Well-being: An Overview)," *Econometric Reviews*, 32, 7–34.

De Vaus, D. A.(2002), *Survey in Social Research*,5th edition, Australia: Routledge

Dobbie, M. J., and Dail, D. (2013), "Robustness and sensitivity of weighting and aggregation in constructing composite indices," *Ecological Indicators*, 29, 270–277.

Donders, A. R. T., Van der Hijden, G. J. M. G., Stijen, T., and Moons, K. O. M., (2006), "Review: A gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, 59, 1087–1091.

Efron, B. and Tibshirani, R. (1993), *An Introduction to Bootstrap*, New York: Chapman and Hall.

Hadi, A. S. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society, Series (B)*, 54, 761–771.

Hadi, A. S. (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," *Journal of the Royal Statistical Society, Series (B)*, 56, 393–396.

Mazziotta, M., and Pareto, A. (2013), "Methods for Constructing Composite Indices: One for all or all for one," *Rivista Italiana Di EconomiaDemogra_a E Statistica*, LXVII, 67-79.

Organization for Economic Co-Operation and Development (OECD), (2008), *Handbook on Constructing Composite Indicators - Methodology and User Guide*, OECD

Penny, K. I., and Jolliffe, I. T., (2002), "A Comparison of Multivariate Outlier Detection Methods for Clinical laboratory," *Journal of Royal Statistical Society*, 50, 295–307.

Rocke, D.M., and Woodruff, D. L., (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, 91, 1047–1061.

Roderick J. A. and Donald B. R., (1987), *Statistical Analysis with Missing*

Data, New York: John Wiley and Sons

Schimitt, P., Mandal, J., and Guedj, M. (2015), "A Comparison of Six Methods for Missing Data Imputation," *Biometrics and Biostatistics*, 6:224. DOI: 10.4172/2155-6180.1000224.

Stephen, P. B. (1998), "Markov Chain Monte Carlo Method and Its Application," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1), 69–100.

Zhang, Z. (2016), "Missing Data Imputation: Focusing on Single Imputation," *Annals of Translation Medicine*, 4(1): 9. DOI: 10.3978/j. ISSN: 2305-5839.2015.12.38.

Zhou, P., Ang, B. W., and Poh, K. L. (2007), "A Mathematical Programming Approach to Constructing Composite Indicators," *Ecological economics*, 62(2), 291–297.

The Change in the Patterns of the Egyptian Family Expenditure in Health (2005-2013)

Amani Musa Mohammed **Shimaa Farouk Hassan***

Abstract

The study of income, expenditure and consumption is considered one of the most important household studies which conducted by statistical agencies in different countries of the world providing a large amount of data upon which we can measure the standard of living for families and individuals as well as establish the rules and information to measure poverty and the tools for weights required for the compilation of Consumer Price Index which is an important indicator to measure inflation.

This paper is concerned with the study of behavior for Health family expenditure in Egypt in rural, urban and total through research income, expenditure and consumption survey in (2004 / 2005), (2008 / 2009), (2010 / 2011) and (2012 / 2013). Pearson curves were used to fit the probability functions to obtain the important statistical measurements necessary to compare their patterns. This paper is organized as follow: in section 2 the Pearson system has been discussed. Sources of data and methodology are introduced in section 3. Finally, Section 4 contains the conclusions.

Keywords: Expenditure, household survey, skewness, moments, Pearson distribution system, kurtosis.

*PhD Student, Department of Applied Statistics and Econometrics , Faculty of Graduate Studies for Statistical Research, Cairo University.

1-Introduction:

The Arab Republic of Egypt has realized the importance of the income research for a long time where the research was conducted first in 1958 / 1959 and was followed by many of the researches. Normally, this survey is conducted every five years and it is metaphorically called the household budget survey since it reflects the household patterns of spending on goods and services. There are several authors studied the statistical properties of household consumption-expenditure. see Sahn and Stifel (2003), Barigozzia et. al (2012), Morad (2012,2013).

2- The Pearson Curves

The Pearson system was originally devised in an effort to model visibly skewed observations. It was well known as how to adjust a theoretical model to fit the first two moments of observed data. Any probability distribution can be extended straightforwardly to form a location-scale family. Except in pathological cases, a location-scale family can be made to fit the observed mean (first moments) and variance (second moments) arbitrarily as well. However, it was not known how to construct probability distributions in which the skewness (standardized third moments) and kurtosis (standardized fourth moments) could be adjusted equally freely. This need became apparent when trying to fit known theoretical models to the observed data that exhibited skewness. Pearson's examples include survival data, which are usually asymmetric. The family of Pearson curves consists of 12 types plus the normal distribution. Many very important distributions in mathematical statistics may be obtained by a transformation. The most important distributions for applications are the Types I, III, VI, and VII. See Pearson (1895, 1901, and 1916).

Later many books explained Pearson curves Elderton and Johnson(1969).

Pearson curves can be used in many fileds of science such as: health, economic and demographic. See Verma and Loh (2008), Pizzutilo (2012) and Stavroyiannisa et. al (2012).

The method of fitting Pearson curves to some empirical distributions is as follows. Independent observations are used to calculate the first four sample moments, and then the suitable type of Pearson curve is determined, using the method of moments to find the unknown parameters.

3- Sources of Data and Methodology

The data used in this paper are obtained from the research of income, expenditure and consumption which conducted by the Central Agency for Public Mobilization and Statistics in the years (2004 - 2005 / 2008 – 2009/ 2010-2011/ 2012 - 2013).

Table1. Frequency Table for Health Annual Expenditure in Urban and Total of the Arab Republic of Egypt in (2004-2005 / 2008-2009)

Health	Urban 2005	Rural 2005	Total 2005	Urban 2009	Rural 2009	Total 2009
Less than 200	5134	10098	15232	1316	2636	3952
200-	7261	9247	16508	2641	4023	6664
500-	4745	3813	8558	2480	2949	5429
1000-	2711	1348	4059	2199	1892	4091
2000-	903	372	1275	1137	703	1840
3500-	311	85	396	438	199	637
5000-	144	47	191	239	99	338
7000-	76	19	95	156	49	205
10000-	43	7	50	75	33	108
More than 15000	39	7	46	53	22	75

Source**: Central Agency for Public Mobilization and Statistics**

Table2. Frequency Table for Health Annual Expenditure in Rural, Urban and Total of the Arab Republic of Egypt

Health	Urban 2011	Rural 2011	Total 2011	Urban 2013	Rural 2013	Total 2013
Less than 200	**236**	**420**	**656**	**167**	**268**	**435**
200-	**566**	**801**	**1367**	**397**	**618**	**1015**
500-	**837**	**980**	**1817**	**634**	**904**	**1538**
1000-	**815**	**1014**	**1829**	**908**	**1099**	**2007**
2000-	**541**	**524**	**1065**	**596**	**717**	**1313**
3500-	**223**	**171**	**394**	**254**	**265**	**519**
5000-	**148**	**96**	**244**	**166**	**166**	**332**
7000-	**97**	**54**	**151**	**105**	**90**	**195**
10000-	**61**	**30**	**91**	**48**	**35**	**83**
More than 15000	**34**	**19**	**53**	**43**	**20**	**63**

in (2010-2011 / 2012-2013)

Source**: Central Agency for Public Mobilization and Statistics**

To find the curves which represent the Health Annual Expenditure of the Rural, Urban and Total of Egypt we computed the values of β_1, β_2 and K.

Where: $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\beta_2 = \frac{\mu_4}{\mu_2^2}$

and $K = \frac{\beta_1(\beta_2+3)^2}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)}$

The K criterion measures the extent of deviation from the symmetrical curve. A negative value of K indicates that the curve in question is

negatively asymmetrical while a positive value of K indicates that the curve is positively asymmetrical. The K criterion may have any value from - ∞ to + ∞, and the different Types of Pearson curves cover all these possible values without overlap.

The programs are written using Maple V Release 4 to compute the values of K, the parameters and the statistical measurements of the models.

Table3. Values of B1, B2 and K for Rural, Urban and Total of the Arab Republic of Egypt

Health	β_1	β_2	K
Urban 2005	41.66775	63.28568	-80.59467
Rural 2005	86.60018	161.18643	26.79974
Total 2005	61.43623	97.42088	166.36963
Urban 2009	1 5.86613	2 4.32382	- 12.03632
Rural 2009	35.33882	55.3343	-193.42688
Total 2009	4.34624	7.64964	-1.87671
Urban 2011	9.13566	14.14132	-4.49126
Rural 2011	15.45571	24.28468	-14.91824
Total 2011	11.89224	18.27031	-7.00153
Urban 2013	8.60817	13.65322	-4.58856
Rural 2013	9.85608	17.00678	-16.49522
Total 2013	9.60800	15.71116	-7.26671

From the above table 3, it is noticed that the first type and the sixth type of Pearson curves can be used. For Rural 2005, Total 2005, the sixth type of Pearson curve is used since K is greater than zero. While for Urban 2005 ,Urban 2009, rural 2009, Total 2009, Urban 2011 , Rural 2011, and Total 2011, Urban 2013 , rural 2013, Total 2013 , the first type of Pearson curve is used since K is less than zero.

The models are fitted as follows:

1-Urban (2004 – 2005):

$$y = y_0 \left(1 + \frac{x}{a_1}\right)^{m_1} (1 - \frac{x}{a_2})^{m_2}$$

$$f(x) = 0.32672375\left(1 + \frac{x}{41346.76539}\right)^{-0.74786136}\left(1 - \frac{x}{-7505.84178}\right)^{4.119677}$$

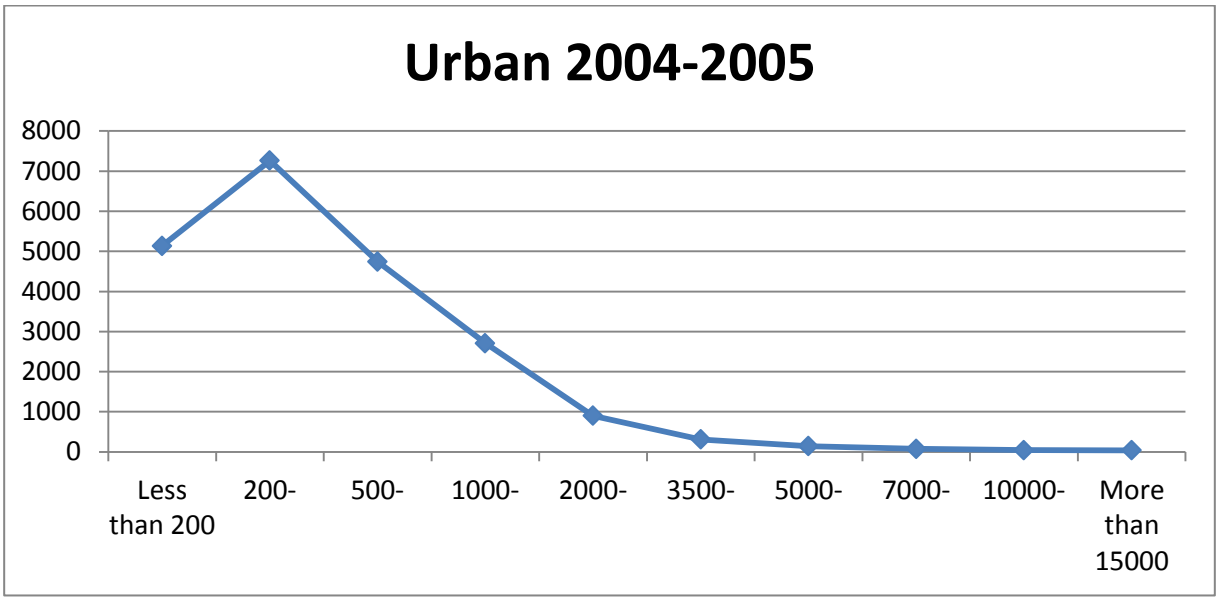


Figure (1): The Curve of Urban in 2004-2005

2-Rural (2004 – 2005):

$$f(x) = 0.6531941 \quad e \quad -35 \quad (x - 18151.26278) - 8.879387 \quad x \quad -0.92509337$$

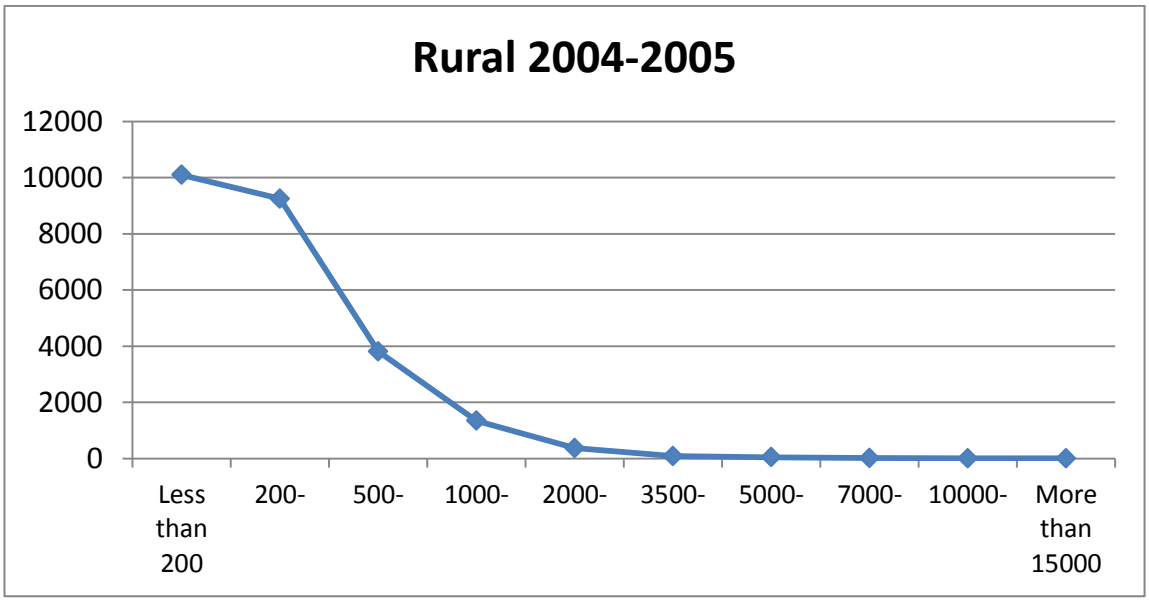


Figure (2): The Curve of Rural in 2004-2005

3-Total (2004 – 2005):

f (x)= **0.1886117391 e -245 (x – 180487.420) -47.37590596 x -0.93010172**

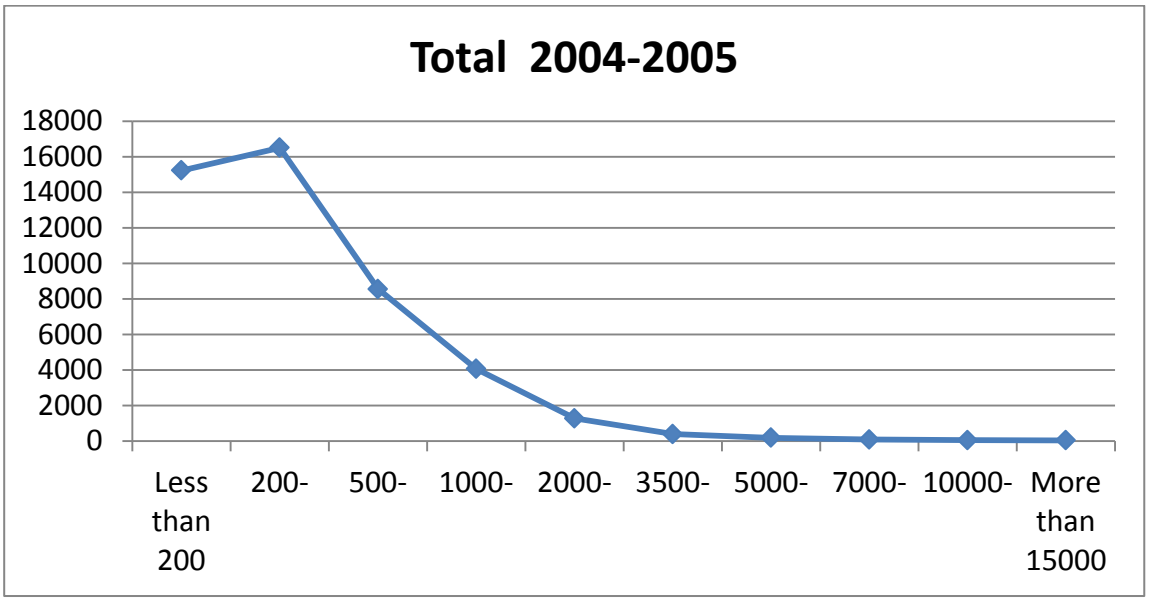


Figure (3): The Curve of Total in 2004-2005

4-Urban (2008 – 2009):

f (x) = -0.4721849000 e -39 (x – 48127.41415) -0.823213737 x 7.861467553

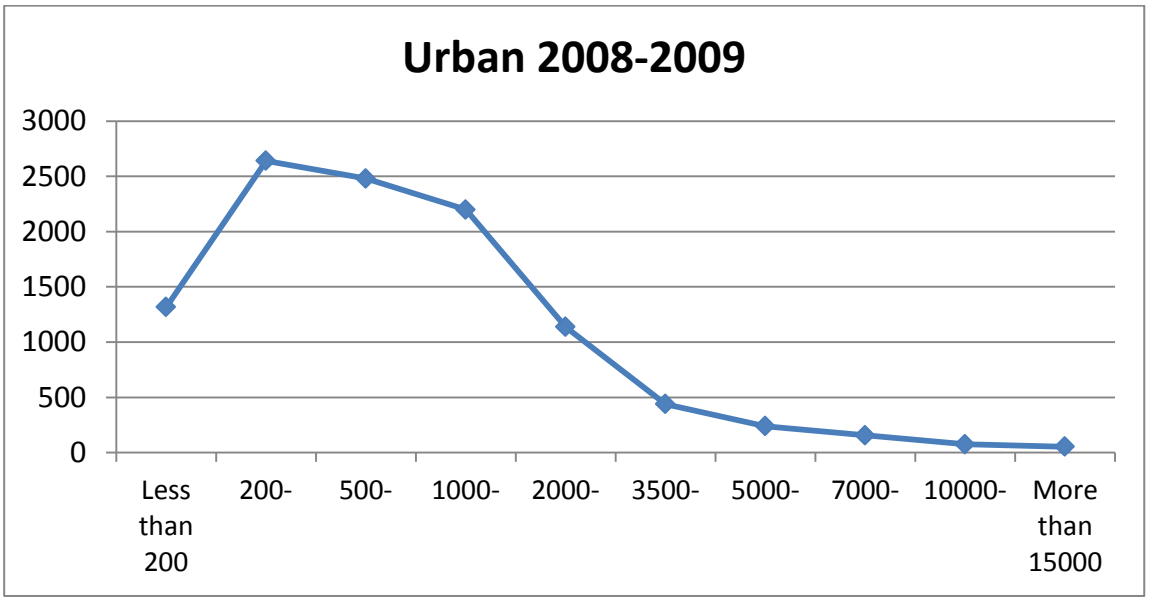


Figure (4): The Curve of Urban in 2008-2009

5-Rural (2008 – 2009):

$$f(x) = -0.2521925\ e^{-465\left(1+\frac{x}{359744.7592}\right)^{-0.891124}\left(1+\frac{x}{-3841.3033}\right)^{83.45537}}$$

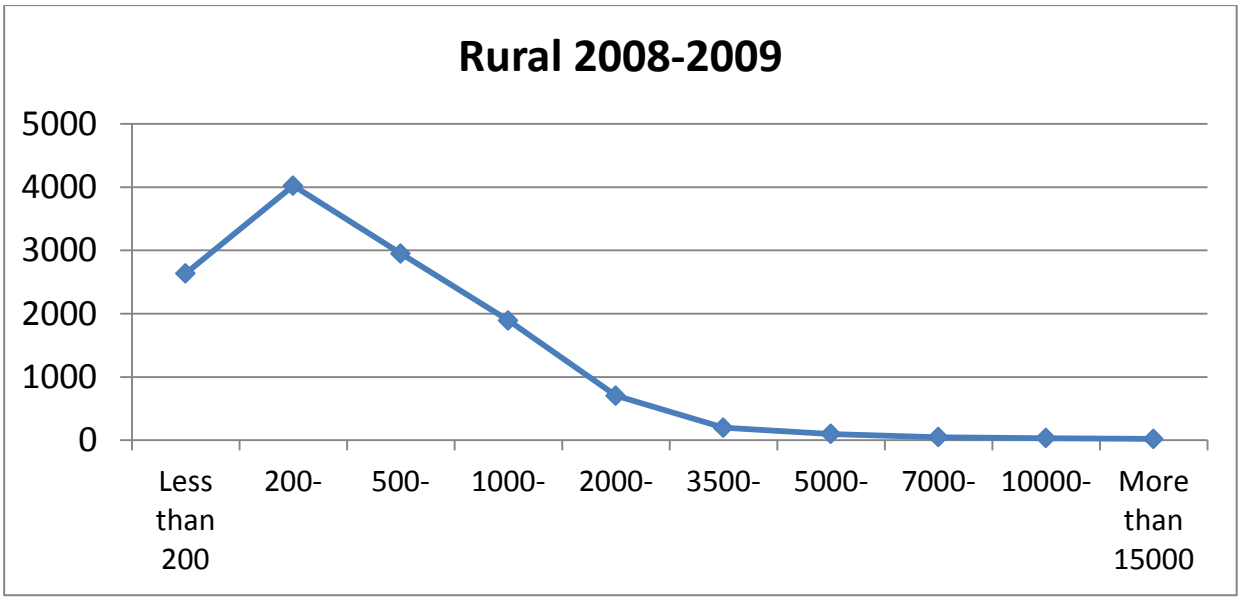


Figure (5): The Curve of Rural in 2008-2009

6-Total (2008 – 2009):

$$f(x) = 3.62672804\left(1+\frac{x}{136513.1630}\right)^{-0.644644}\left(1+\frac{x}{-37599.76795}\right)^{2.340505}$$

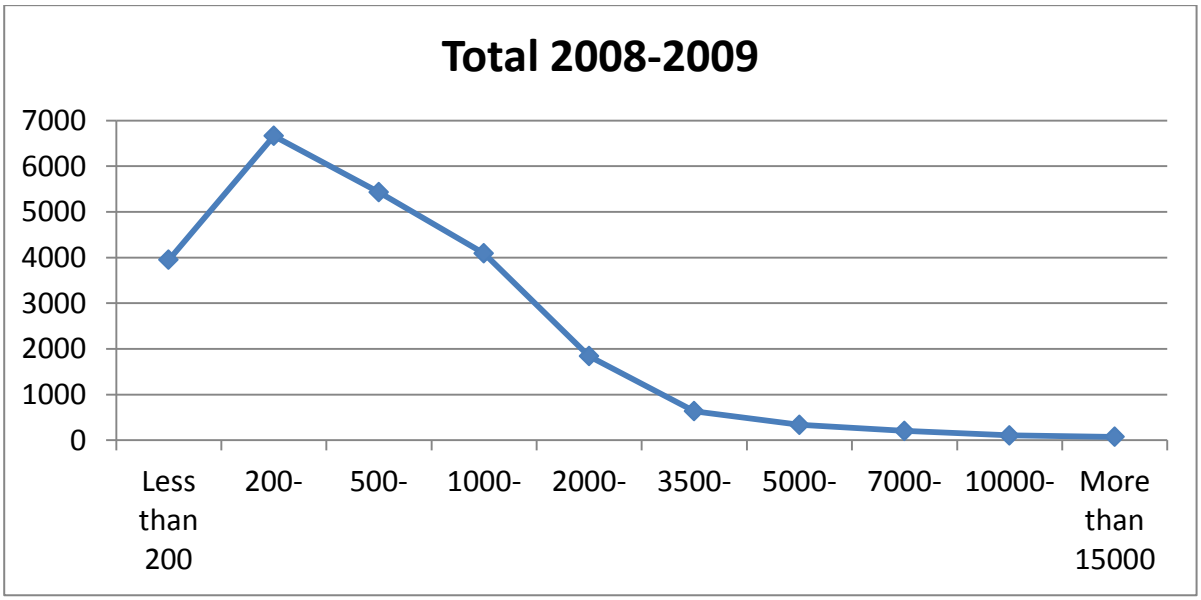


Figure (6): The Curve of Total in 2008-2009

7-Urban (2010 – 2011):

$$f(x) = 0.35029206 \left(1 + \frac{x}{39375.64886}\right)^{-0.7757497} \left(1 - \frac{x}{-8813.17655}\right)^{3.465906}$$

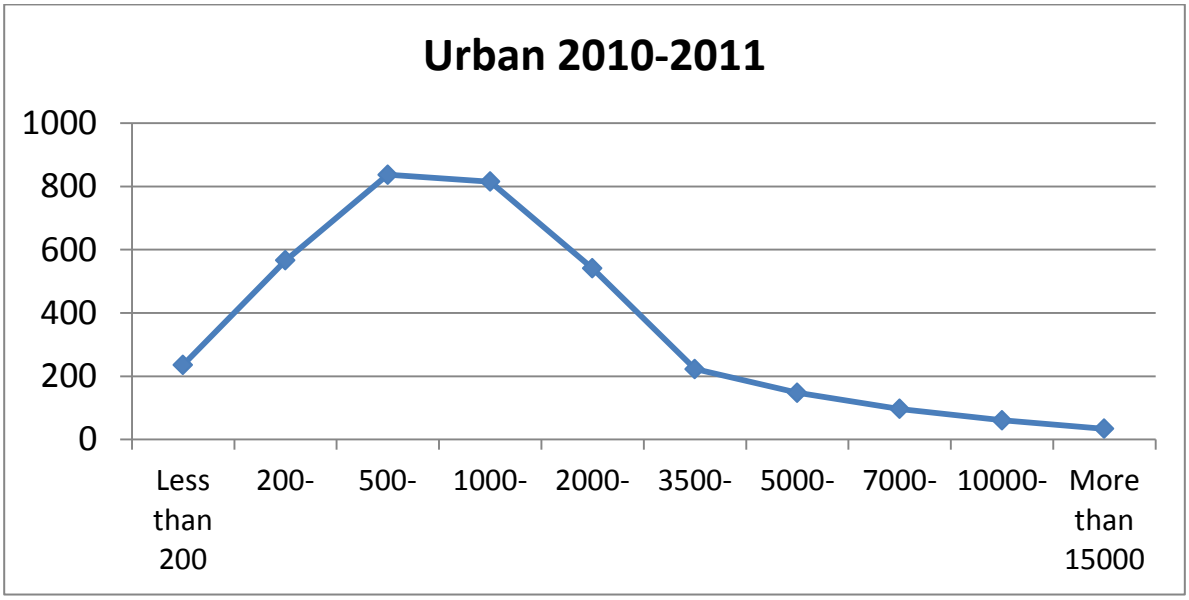


Figure (7): The Curve of Urban in 2010-2011

8-Rural (2010 – 2011):

$$f(x) = 0.5510898 \left(1 + \frac{x}{64950.74804}\right)^{-0.802594} \left(1 - \frac{x}{-4666.30064}\right)^{11.171397}$$

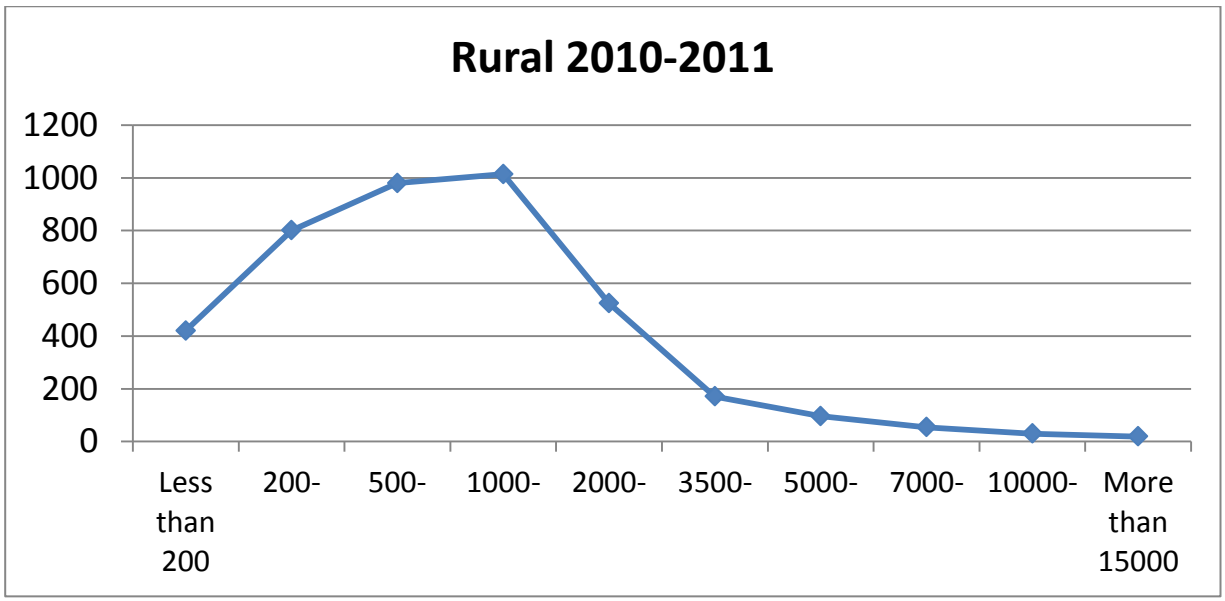


Figure (8): The Curve of Rural in 2010-2011

9-Total (2010 – 2011):

$$f(x) = 0.85479913 \left(1 + \frac{x}{43729.40078}\right)^{-0.7971557} \left(1 - \frac{x}{-6862.30350}\right)^{5.0798017}$$

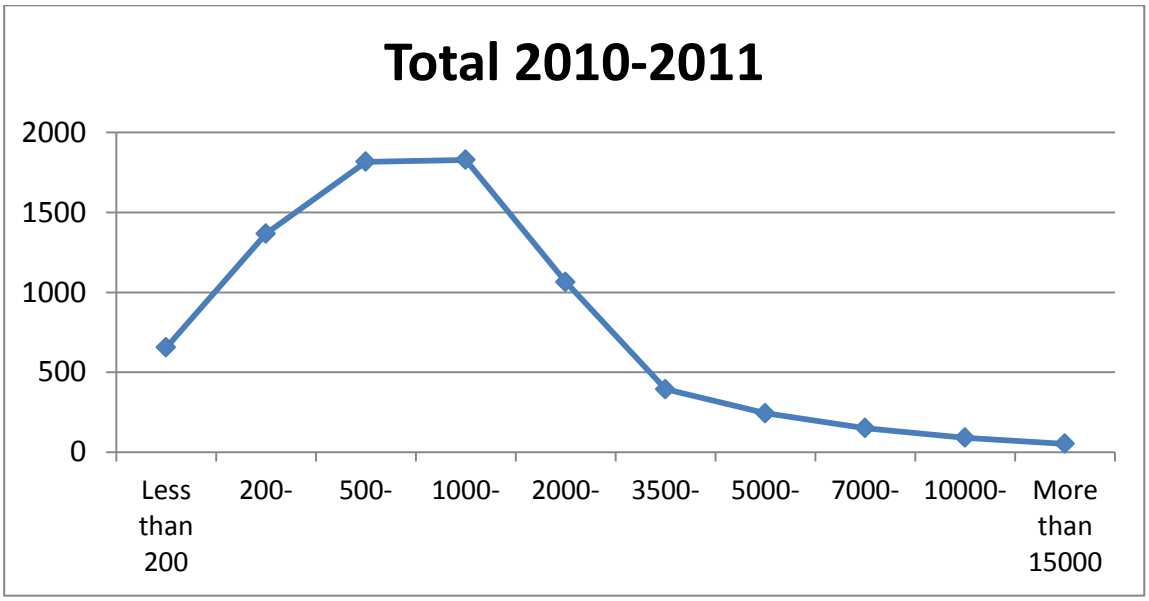


Figure (9): The Curve of Total in 2010-2011

10-Urban (2012 – 2013):

$$f(x) = 0.32672375 \left(1 + \frac{x}{41346.76539}\right)^{-0.74786136} \left(1 - \frac{x}{-7505.84178}\right)^{4.119677}$$

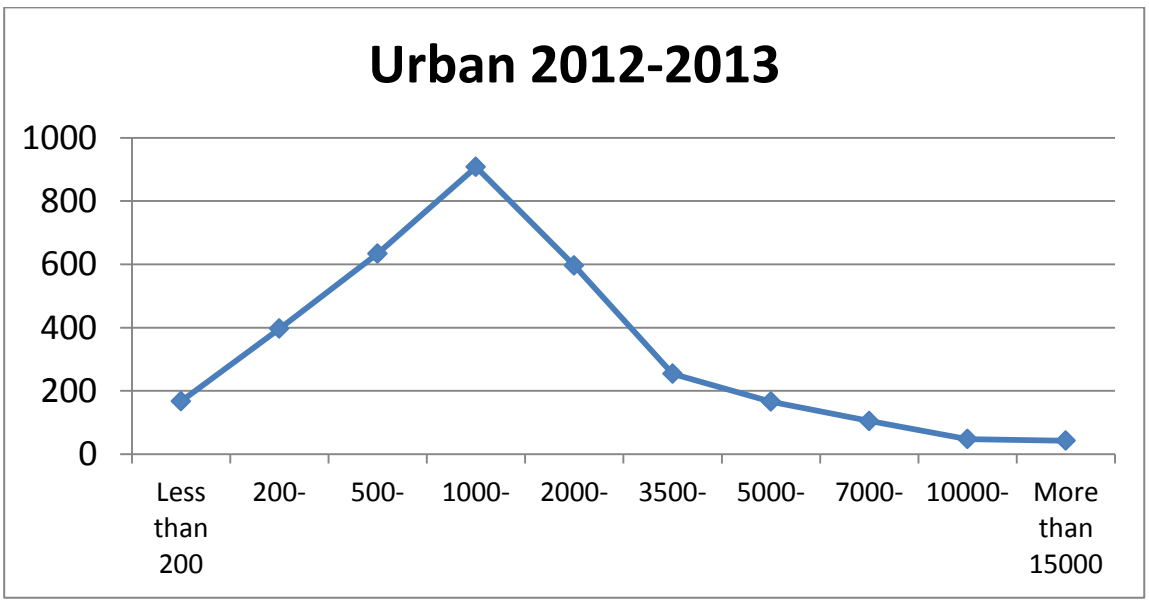


Figure (10): The Curve of Urban in 2012-2013

11-Rural (2012 – 2013):

$$f(x) = 0.58841349 \left(1 + \frac{x}{96390.82373}\right)^{-0.655812} \left(1 - \frac{x}{-2822.93473}\right)^{22.393113}$$

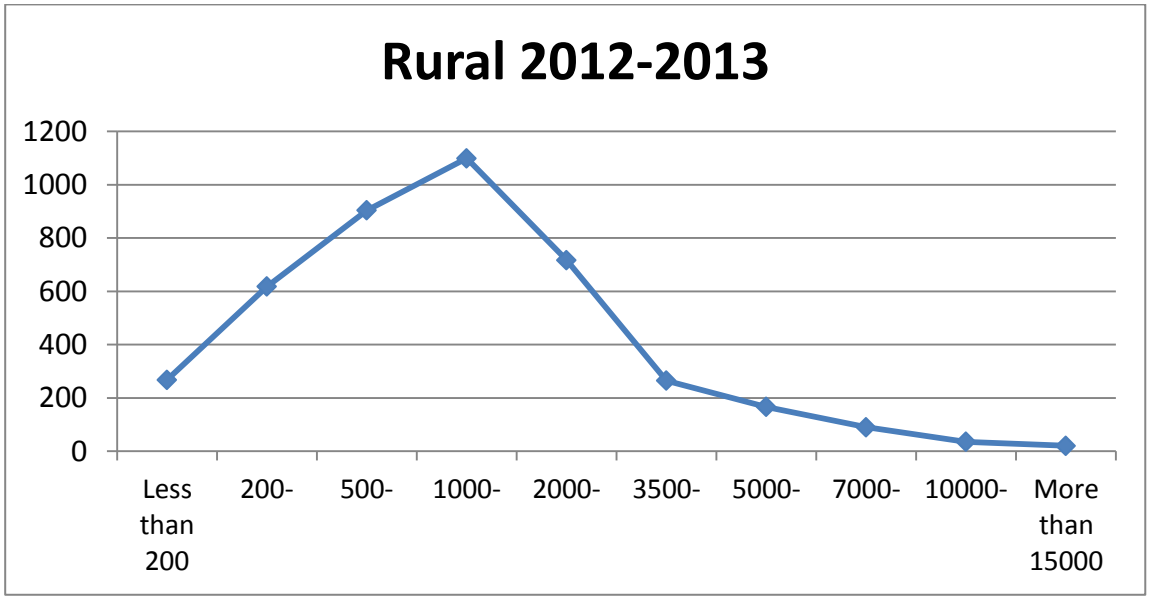


Figure (11): The Curve of Rural in 2012-2013

12-Total (2012 – 2013):

$$f(x) = 0.87313909 \left(1 + \frac{x}{50820.78565}\right)^{-0.7190194} \left(1 - \frac{x}{-4733.23207}\right)^{7.720122}$$

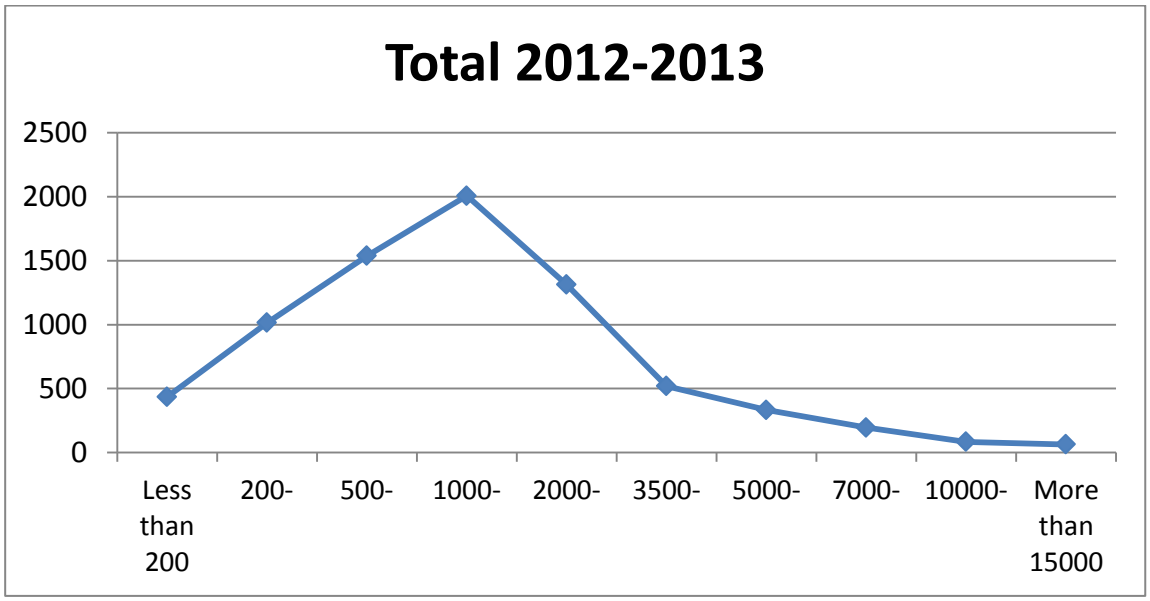


Figure (12): The Curve of Total in 2012-2013

To understand more about the Pearson curves, and to compare among rural, urban and total of the Arab Republic of Egypt from (2004 – 2005) to (2012 - 2013), some statistical measurements are needed to compute.

Measurements of central tendency such as the mean, the median and the mode are computed.

Table 4. Values of Statistical Measurements for Rural, Urban and Total of the Arab Republic of Egypt

	Mean	Median	S.D	C.V	Skewness	Kurtosis
Urban 2005	**466.024**	**278.625**	**684.970**	**146.982**	**9.306**	**161.186**
Rural 2005	**817.695**	**429.286**	**1320.732**	**161.519**	**6.537**	**65.336**
Total 2005	**1480.669**	**784.274**	**2103.565**	**142.068**	**3.983**	**24.323**
Urban 2009	**627.933**	**344.893**	**1042.584**	**166.034**	**7.838**	**97.421**
Rural 2009	**907.592**	**473.415**	**1379.678**	**152.015**	**5.945**	**55.334**
Total 2009	**1171.159**	**597.025**	**1773.347**	**151.418**	**4.7635**	**34.497**
Urban 2011	**907.592**	**473.415**	**1379.678**	**152.015**	**5.945**	**55.334**
Rural 2011	**2133.010**	**1171.779**	**2733.760**	**128.164**	**3.022**	**14.141**
Total 2011	**2399.337**	**1507.709**	**2835.505**	**118.179**	**2.934**	**13.653**
Urban 2013	**1838.900**	**998.211**	**2414.782**	**131.317**	**3.448**	**18.270**
Rural 2013	**1967.838**	**1273.885**	**2248.852**	**114.280**	**3.139**	**17.007**
Total 2013	**2158.733**	**1379.671**	**2534.332**	**117.399**	**3.099**	**15.711**

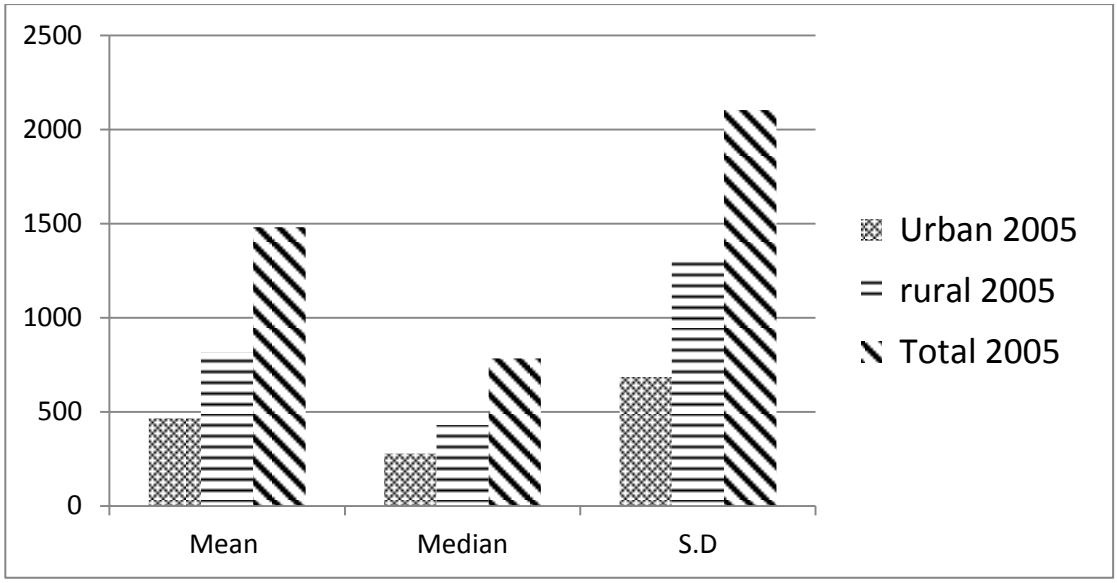


Figure (13): The Values of Mean, Median, Standard Deviation for Three Places in 2004-2005

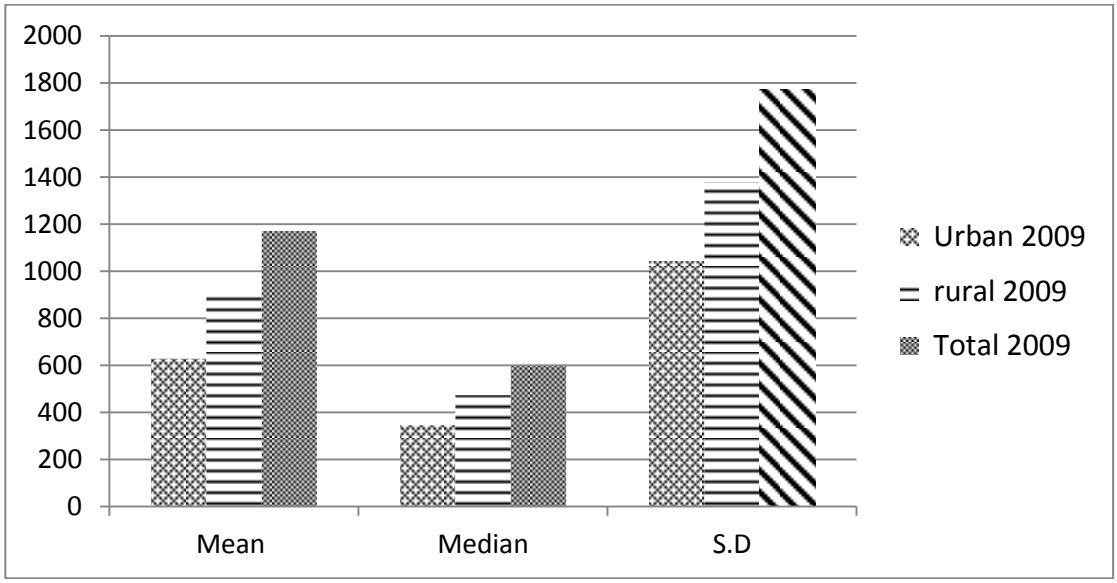


Figure (14): The Values of Mean, Median, Standard Deviation for Three Places in 2008-2009

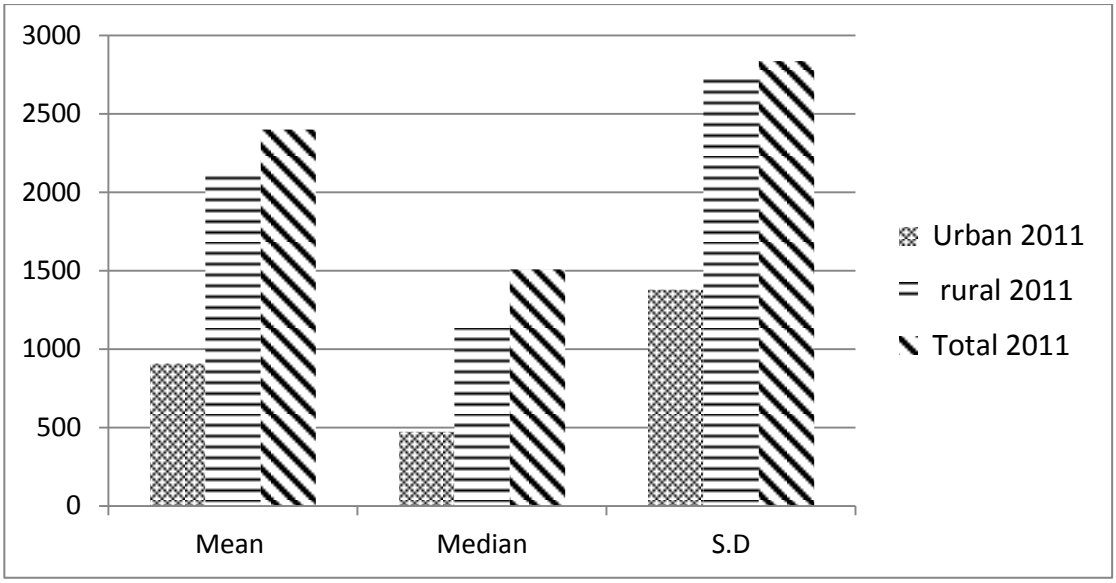


Figure (15): The Values of Mean, Median, Standard Deviation for Three Places in 2010-2011

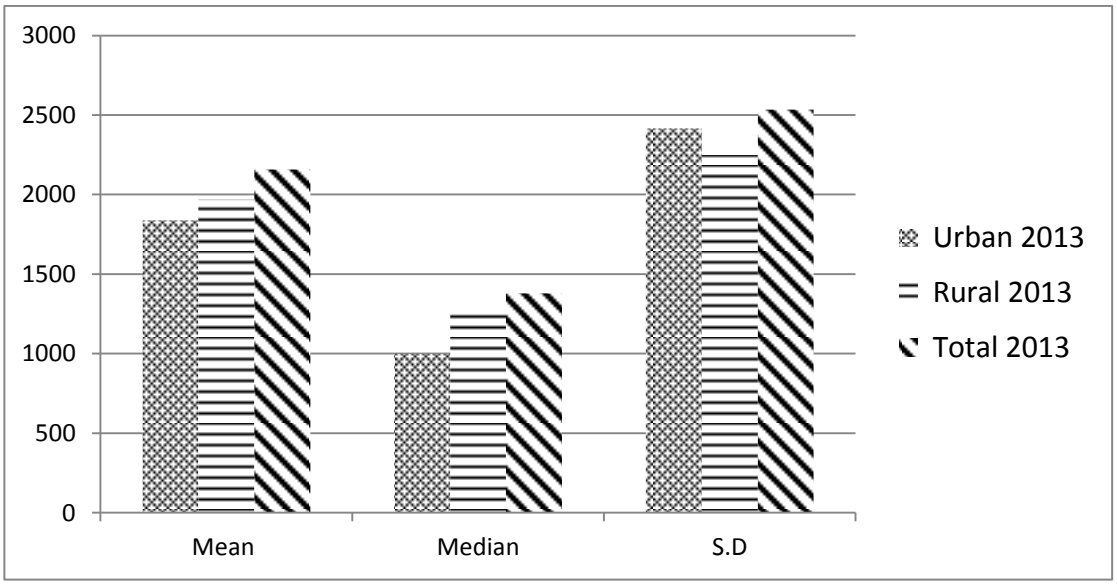


Figure (16): The Values of Mean, Median, Standard Deviation for Three places in 2012-2013

4-Conclusion:

From the above models, figures (from 1 to 16) and values of tables 4; It may be noted that there is a difference in the pattern of the expenditure on Health between rural and urban areas and the total of the republic of (2004/2005) to (2012/2013) ,we note that the spending decreased on Health (2004/2005) to (2012/2013) also found that the values of the average , median , mean and standard deviation of the rural are greater than urban and total of the Republic.

The mean is trend to increase from 2005 to 2013 for the urban and rural. Always the mean of rural is greater than urban. On the other hand, we notice that C.V is trend to decrease. We notice that skewness for 2005 and 2009 has positive value it means that it skews to the right.

The values of the kurtosis for rural, urban areas and the total from (2004/2005) to (2012/2013), except for Total 2011, are greater than 3 so they are sharper than a normal.

References

Barigozzia M., Alessib L., Capassoc M., Fagiolod G. (2012), "The distribution of household consumption-expenditure budget shares", Structural Change and Economic Dynamics, Volume 23, Issue 1, PP. 69–91

Elderton, W. P., and Johnson, N. L. (1969), *the Systems of Frequency Curves* , the Syndices of the Cambridge University Press.

Kendall, A.S.,(1958) *The Advanced Theory of Statistics* (VOL . 1) , London.

Mousa , A. (1974) A statistics model for consumption in A . R. E, M S C thesis submitted to the Department of Statistics, Faculty of Economics and political science, Cairo University.

Morad, A.Naglaa (2013). "Family Pattern of Expenditure in Egypt (2005-2010) ". The 38th International Conference For Statistics, Computer Sciences and Its Application

Pearson, Karl (1893). "Contributions to the mathematical theory of evolution". Proceedings of the Royal Society of London 54, PP. 329–333.

Pearson, Karl (1895). "Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material". Philosophical Transactions of the Royal Society of London.

Pearson, Karl (1901). "Mathematical contributions to the theory of evolution X: Supplement to a memoir on skew variation". Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 197, PP. 443–459.

Pearson, Karl (1916). "Mathematical contributions to the theory of evolution, XIX: Second supplement to a memoir on skew variation". Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 216, PP. 429–457.

Pizzutilo,F.,(2012)"Use of the Pearson System of Frequency Curves for the Analysis of Stock Return Distributions: Evidence and Implications for the Italian Market", Economics Bulletin, Vol. 32, No. 1, pp. 272-281, 2011

Sahn,D.E. and Stifel,D (2003). "Exploring Alternative Measures of Welfare in the Absence of Expenditure Data". Review of Income and Wealth.Volume 49, Issue 4, PP. 463–489,

Stavroyiannisa, S., Makrisa, I. Nikolaidisa, V. Zarangasb, L.(2012) "Econometric modeling and value-at-risk using the Pearson type-IV distribution" International Review of Financial Analysis , V.22, PP. 10–17

Verma, R.B.P. and Loh, S.,(2008)," An Evaluation of the Pearsonian Type I Curve of Fertility for Aboriginal Populations in Canada, 1996 to 2001", Canadian Studies in Population, Vol. 35.2, 2008, pp. 357–372.

- بحوث الدخل والإنفاق لسنوات 2005/2004 ، 2009/2008 و2011/2010 و2012 / 2013 الجهاز المركزي للتعبئة العامة والإحصاء .

النظرية الاحصائية وتطبيقاتها (1969) – محرم وهبي محرم – القاهرة .

An Econometric Analysis of Modernized Agriculture in Reducing Multidimensional Poverty in Uganda.

Emong Herbert Robert

E-mails:herbertist64@yahoo.co.uk

Abstract:

This paper is aimed at developing an econometric analysis on how modern agriculture can be a fundamental instrument for reducing multidimensional poverty in Uganda. It demonstrates the importance of agriculture in reducing inequalities amongst the poor while focusing on the relationship between increasing productions from modern agricultural practices and poverty level across the country. The study utilizes the applications of Box-Jenkins approach to cereal production data and econometric analysis as the main tool to understand the underlying dynamics of modern agricultural practices in Uganda.

Keywords: Agriculture,GDP,Growth,Household,Livelihood,Modernization,Multidimensional Poverty, Uganda.

1. Introduction.

Uganda`s dependence on agriculture and her aspirations for modern agriculture dates back to the colonial times. Over the years, the common theme has been to commercialize the agricultural sector described by Flygare (2006) as subsistence-based. This objective for long has been referred by Nabudere (1997) as the politics of modernization in Africa. Major researches have been carried on how Uganda can benefit from modernization of agriculture acknowledging its importance to the economy and lots of suggestions have been for the whole concept of modernization practiced as universal and desirable to be revisited.

The starting point being the actual experience of the poor who are adversely affected by these strategies if meaningful transformation is to be achieved (Wiggin, 2010).

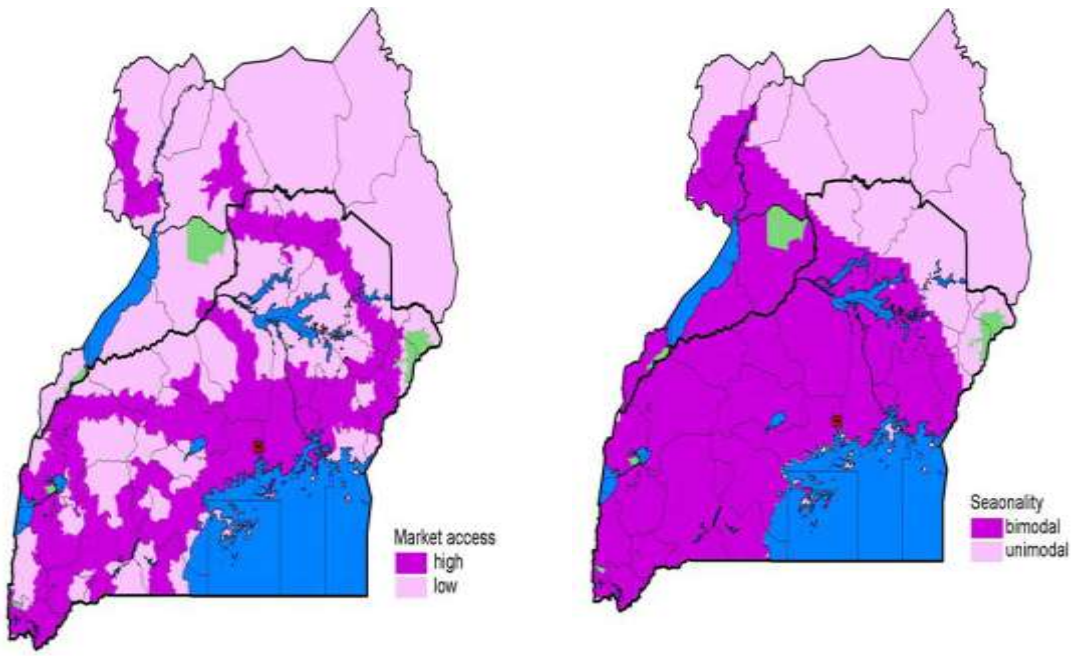
Most foods consumed in Uganda are produced by the local subsistence agricultural system and contributes the most revenue earnings. The transformation is therefore expected to be a modern system where farm productivity is high due to the employment of modern agricultural technologies and incomes and employment levels both rural and urban will improve. These expectations have been pursued through national government`s driven agriculture modernization strategy in conjunction with external agencies that are working towards determining ways in which food production and incomes can be sustained (Flygare ,2006).

Following this introduction, this paper is structured as follows: Section 1 provides an introduction of modern agriculture in Uganda. The literature review inculuding the agricultural potentials and the roles played are described in section 2. Multidimensional poverty in Uganda,approaches and some common measures of poverty in section 3. Section 4 provides the main results and discussions using the Minitab and SPSS programs. The agricultural share to GDP and conclusion in section 5 and 6 respectively.

2. Literature Review

The great diversity of agro-ecological zones in Uganda determine the natural agricultural potentials and analysis shows that Uganda is enriched by spatial differentiations. They cover about three quarters of the country with most of these zones receiving bimodal rainfall that allows double harvest with the application of improved agricultural practices. This has further caused differences in income generation and the impacts realized from modern agriculture to poverty across all the regions in Uganda due to differences within yields, investments in infrastructure and greater access to local and national markets as shown on Figure 1 (i & ii) respectively (Ruecker,2003).

Figure 1: Domains that Determine Natural Agricultural Potential and Market Access in Uganda.



(i) **(ii)**

Source: Ruecker *et al*. (2003).

Half of the country has bimodal rainfall distribution, allowing double harvests in those areas. Most of these territories have the favorable length of growing period with exception of the north-east Uganda.The population density in those areas grows not only due to the natural birth but also migration of people from other parts of the country. Market access is better within these areas due to improved infrastructural developments and the actual farm performance differ by regions coinciding with these agricultural potential zones. Areas with favorable conditions like good climates, better infrastructural developments and better yields of the major crops tend to have better access to markets with these crops fetching pleasant prices than in other regions (USAID ,2007).

Almost all agricultural productions take place on small plots of lands with about 40% of the plots mixed with lots of other food crops. The sector however registered a positive growth rates of 2.6% in 2009 thou this is below the targeted rate of 6 % per annum as set by the African Union. Uganda is not alone with its small farms,in most countries, both rich and poor, the average size of farms is often small and the sector is dominated by the small owner-operated family units that combine ownership of the main factors of production (UBOS, 2014).

Farms in Asia are actually smaller than in Sub-Saharan Africa on average, they are not very big in Central America and Europe yet this has not prevented those small farms to generate a large pro-poor growth. Eastwood (2010) emphasized on a systematic inverse relationship between farm size and crop income per acre. Smaller farms produce the largest return per acre which is 3 times bigger than the larger farms at a national level.

Size of farms is largely determined by the land availability. Large farms are rare phenomenon in the Uganda with land constraints and high population density. The land endowment in Uganda is comparable to Malawi, but its other neighbors Kenya, Tanzania, Democratic Republic of Congo and Sudan enjoy 2-3 times higher rates. It therefore seems that the most viable strategy for agricultural development in Uganda is a smallholder strategy with the gradual changes in farm size, a subject to household endowment, preferences and rural outmigration

(Fuglie, 2009).

Table 1: Land Area in Selected Countries, Hectare (ha) Per Capita.

Selected Countries	1960-70	1971-80	1981-90	1991-00	2001-08
Ethiopia	3.91	3.04	2.39	1.76	1.35
Kenya	5.93	4.11	2.83	2.05	1.62
Malawi	2.36	1.74	1.22	0.90	0.72
Sudan	17.91	13.31	9.79	7.59	6.19
Tanzania	7.54	5.44	3.99	2.92	2.30
Uganda	2.45	1.77	1.30	0.92	0.70
Zambia	20.90	14.87	10.76	8.02	6.39
Sun-Sahara Africa	9.24	6.97	5.23	3.90	3.14
World	3.89	3.16	2.65	2.27	2.02

Source: World Bank Development Indicators (2009).

It is true that larger farms tend to be more commercial thou the sector in general facing challenges of limited availability of farm lands, poor infrastructural developments, lack of capital, low production and productivity, poor marketing system, human resource constraints and over reliance on unpredictable weather conditions. Major transformation aimed at liberalizing agriculture markets, reducing trade barriers and promoting traditional and non-traditional exports may therefore be an ultimate tool for reviving the economy. It is expected to improve rural food supplies, incomes, propel national development and reduce multidimensional poverty across Uganda (World Bank ,2009).

2 .1 The Role of Agriculture.

Agriculture has been and continues to be the most important sector in Uganda's economy employing the largest proportion of the population aged 10 years and older. Major roles have been played by the sector in Ugandan economy including employment opportunities, rural household incomes, food supplies and multidimensional poverty reduction. However, its share has significantly declined over time accounting for 56% of GDP in early 80`s but a recent study found the share to have fallen to 15.4% (Table.3). This declining share has not only affected livelihoods and food security in Uganda but also raised major concerns in other sub-Saharan African countries that highly depend on importation of agricultural products from Uganda. There is therefore need to increase agricultural productivity through modern agricultural practices (World Bank, 2012).

Table 2 : Share of Primary Growth Sectors in GDP and Growth Performance in Uganda.

Growth Sectors In GDP	% share in GDP					% annual growth				
	1988	1997	2004	2007	2008	1988-97	1998-02	2004-08	2007	2008
Agriculture	51.1	33.1	17.3	14.5	15.4	3.9	5.4	1.1	1.7	2.2
Forestry	2.2	1.7	3.3	3.5	3.4	4.7	7.0	3.9	2.2	4.2
Manufacturing	5.9	8.4	7.0	6.9	7.2	13.2	7.2	6.3	7.6	6.7
Hotels and restaurants	1.1	1.9	4.0	4.1	4.1	13.1	3.8	9.6	9.2	12.5
Mining	0.1	0.6	0.3	0.3	0.4	34.6	8.0	13.0	5.0	10.4
Post and telecom	0.2	0.6	2.0	3.0	3.4	10.1	22.8	26.2	16.1	39.6
Construction	4.1	6.5	11.9	12.2	12.2	6.5	6.3	6.3	4.8	5.8

Source: Republic of Uganda ,2010.

The sectors of tourism, industry and post and telecommunications have double digit growth performance but with very low impacts on the GDP. The manufacturing sector`s contribution to GDP has barely changed since 1988 even with market liberalization and enabling policies, the sector still contributes the same as more than two decades ago. Liberalization of the commodity and finance markets have allowed for uncontrolled flow of capital, since majority investments are foreign investments, and it is clear from the figures that the country experiences a huge capital and resource extraction through repatriation of profits since the biggest portion of surpluses generated from these sectors do not enter into the national economy.

The future of Uganda's agriculture will depend on how the country takes advantage of what science has to offer. Urban and subsistence agriculture have made a significant contribution to the survival of the urban poor. It supplements their incomes, creates employment opportunities and contributes to their food and nutrition security.

3. Multidimensional Poverty in Uganda

The concept of poverty is broader than signifying a mere lack of material resources. It also includes lack of other human needs like freedom, access to clean drinking water, access to power and lack of opportunities and choices. Accordingly, poverty cannot be explained or defined by a single measure. There are a number of general questions about how to define and measure poverty which apply to all approaches.

Alkire and Foster (2010) elaborated on how poverty maps for Uganda and other developing countries are not homogenous and vary widely across space. Some of these differences are caused by differences in agricultural zones and agro-climatic conditions, infrastructural access to markets, water bodies, and political factors. Uganda`s progress in reducing income poverty is strongly reflected in other dimensions of welfare such as education, health, housing conditions and access to information. This multidimensional approach sheds light on a variety of deprivations measured using different indicators and the methodology proposed by Foster.

A household is identified as multi dimensionally poor if the weighted indicators in which they are deprived sum up to 30%. The current study depends on income poverty approach and in particular relative poverty where the poorest ,20% of the population is considered poor (Alkire and Santos, 2010).

Appleton (2001) acknowledged the rapid reduction in poverty in Uganda reflected in the expansion of the middle class, but a large group of insecure individuals has persisting throughout the period as elaborated on Table .4. The growing middle class represents an engine for socioeconomic transformation because of its greater spending power and, more importantly, the ability to save and invest in the future. Nonetheless, a large section of the population who are either poor or insecure non-poor are vulnerable. The dominant insecure non-poor group also require targeted attention as they remain highly vulnerable and are at a risk of falling back into poverty.

Table 3: Trends in Uganda Poverty Status at a National Level.

Years	Population (Millions)	Poor		Insecure non-poor		Middle class	
		Millions	%	Millions	%	Millions	%
1992/3	17.4	9.8	56.4	5.8	33.4	1.8	10.2
1999/00	21.4	7.2	33.8	9.4	43.9	4.8	22.4
2002/03	25.3	9.8	38.8	10.1	39.9	5.4	21.2
2005/06	27.2	8.4	31.1	10.9	40.2	7.8	28.7
2009/10	30.7	7.5	24.5	13.2	42.9	10.0	32.6
2012/13	34.1	6.7	19.7	14.7	43.3	12.6	37.0

Source: Republic of Uganda, 2013.

A number of Ugandans in the middle class has increased by seven times from 1.8 million in 1992/93 to 12.6 million in 2012/13. About 2.6 million Ugandans have acquired middle class status in the last three years (2009/10 – 2012/13) alone. On the other hand, many Ugandans that have moved out of poverty have failed to attain middle-class status. In 2012/13, the number of insecure non-poor individuals (14.7 million) was 2.5 times higher than what it was in 1992/93 (5.8 million). Although this category is classified as non-poor, they are highly vulnerable, and the occurrence of a shock such as a drought can push them into poverty. This explains how importantly the country needs to embark on modern agricultural practices in order to facilitate their initiatives to eradicate poverty within its boundaries (Republic of Uganda, 2013).

4. The Main Results and Discussion

This section explores the behavior of poverty level using modern agriculture as an indicator, interest is focused on increasing productions of major crops in Uganda and its relationship with the poverty level arising from improved agricultural practices. A necessary characteristic of these variables is that they could be aggregated meaningfully and display variation across the regions at the country level. This analysis will assess how modernization of agriculture has improved total production of these major crops, mainly cereals and crop productions hence earning more revenues to the farmers engaged in agriculture.

4.1 Methodology and Data

A statistical analysis in this section is based on the secondary data generated by the Uganda Bureau of Statistics (UBOS) and Uganda Census Agriculture for the period of 1975-2014, a period when Uganda is considered to have registered an average growth rates of about 6.4% per annum. This represent about 50% of the original data, yet according to UBOS, this portion of the data is fairly representative and provide the same findings as the full data.

Applications of the Box-Jenkins approach is explored to cereal production data, in particular sampling at different time intervals in order to determine if there are systematic patterns in cereal production.

Table 5 below presents the descriptive statistics for some selected variables of agriculture for the comparison of their means with some as explanatory variables of the estimated models we present further on.

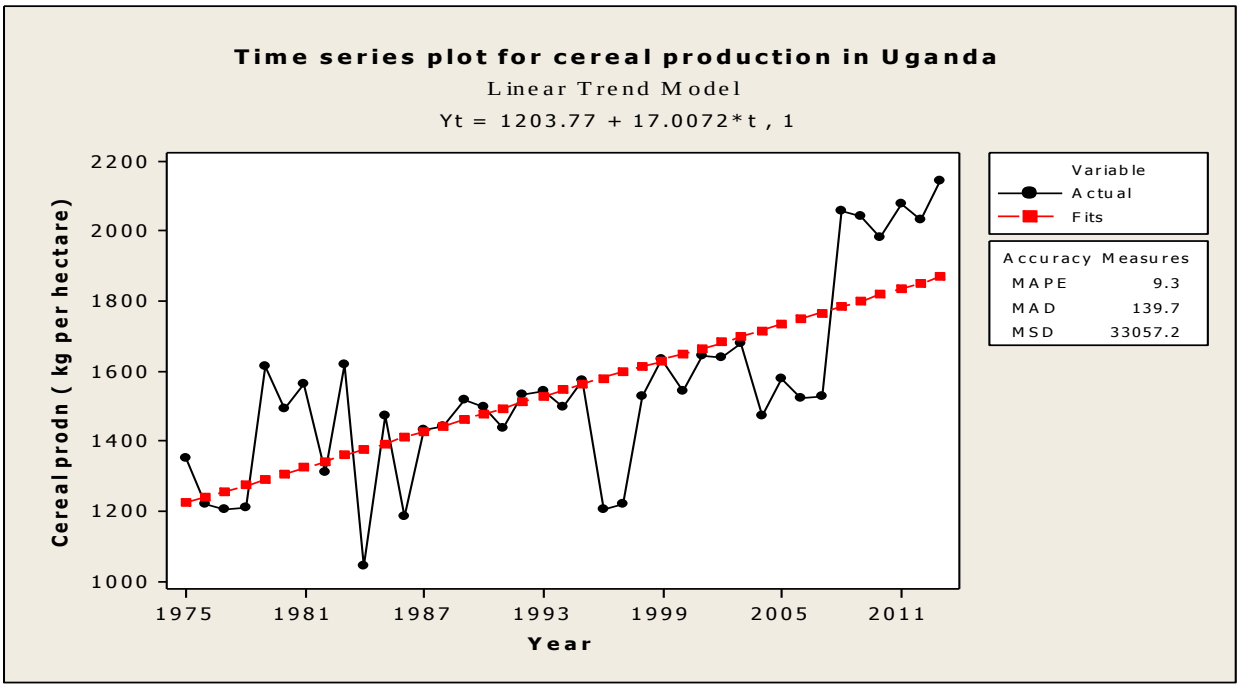
Table 4: Statistical Measures for Some Selected Variables of Agriculture in Uganda.

Statistical Measures	Cereal production	Crop production	Food production	Livestock production
Minimum	1041.9	48.38	45.08	36.44
Maximum	2143.3	111.30	113.58	127.34
Mean	1543.9	81.34	77.06	67.98
Median	1526.0	77.48	70.59	58.58
Variance	71528.9	365.93	468.01	881.11
Standard Deviation	267.4	19.13	21.63	29.89
Coefficient of Variation	17.32	23.52	28.07	43.66
Skewness	0.64	0.02	0.36	0.82
Kurtosis	0.22	1.28	1.27	0.80

The minimum productions for each output were above 35tonnes per hectare with maximums of above 100 tonnes per hectare each indicating more benefits as a result of employing better agricultural practices. Cereal productions are seen to have the lowest relative variations within the major agricultural outputs as a result of more improved and mechanized agricultural practices in Uganda compared to crop, food and livestock productions. This has risen because of the better revenues the rural population generate from cereals in relation to the application of modern agricultural practices hence improving the welfare of more livelihoods.

Trend analysis on Figure. 2 further enhanced spotting of any systematic patterns over time for cereal productions in relations to the application of modern agricultural practices which should be reflected on the poverty levels of the country.

Figure 2: Trend Analysis for Cereal Production (kg per hectare), Model 1

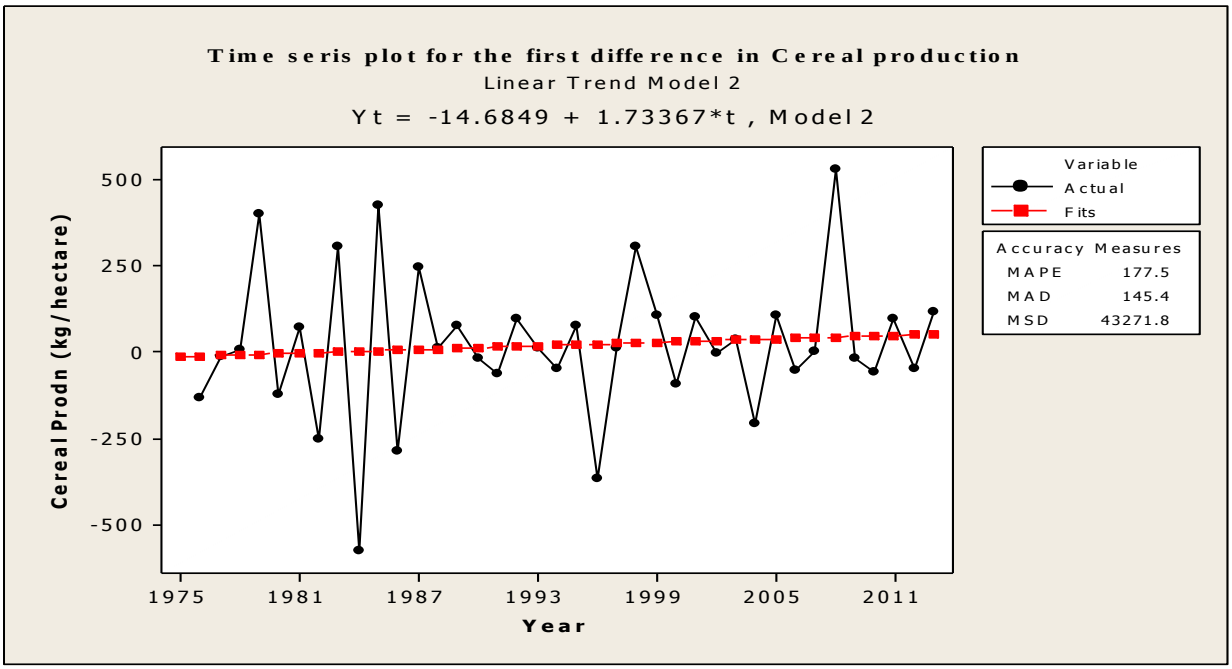


It is apparent that cereal production is having an increasing linear trend and seasonal fluctuations because of the increasing mean and pattern of fluctuations.

In particular, the seasonal variation is in a yearly basis. It can also be noticed that, there are some outliers between the years 1981 to 1987 and after 2005, this is a direct evidence of high agricultural productivities in the early 80s with growing contributions to the GDP from agricultural sector, this is a matter for poverty reduction. An increase in the cereal productions definitely increase the export market share participation and this in return improves the general welfare of the local farmers who are engaged in modern agricultural activities, a positive trend on multidimensional poverty within the regions of Uganda.

General agricultural productivity is seen to have remarkable performance after 2005 because the government and its development partners intensified their efforts toward modern agriculture as a tool to reducing multidimensional poverty in Uganda. The trend spotted in cereal productivities was therefore removed by taking the first difference as shown in figure 3.

Figure 3: First Difference Plot for Cereal Production



The difference series appears to be stationary with presence of some outliers indicating major productions in some years and regions due to the application of modern agricultural technologies within the four regions and therefore a differences in the impacts realized from reduced poverty especially the western side of Uganda which has favorable conditions ranging from improved infrastructural developments to better access to local markets.

Table 5: The Computed RMSE, MAE, MAPE, and R-Squared Values for Model 1

Model	RMSE	MAE	MAPE	R-Squared
ARIMA (1,1,0)	184.616	122.303	8.149	0.606
ARIMA (1,1,1)	193.41	127.484	8.661	0.567
ARIMA (1,1,2)	145.6	85.639	6.035	0.778

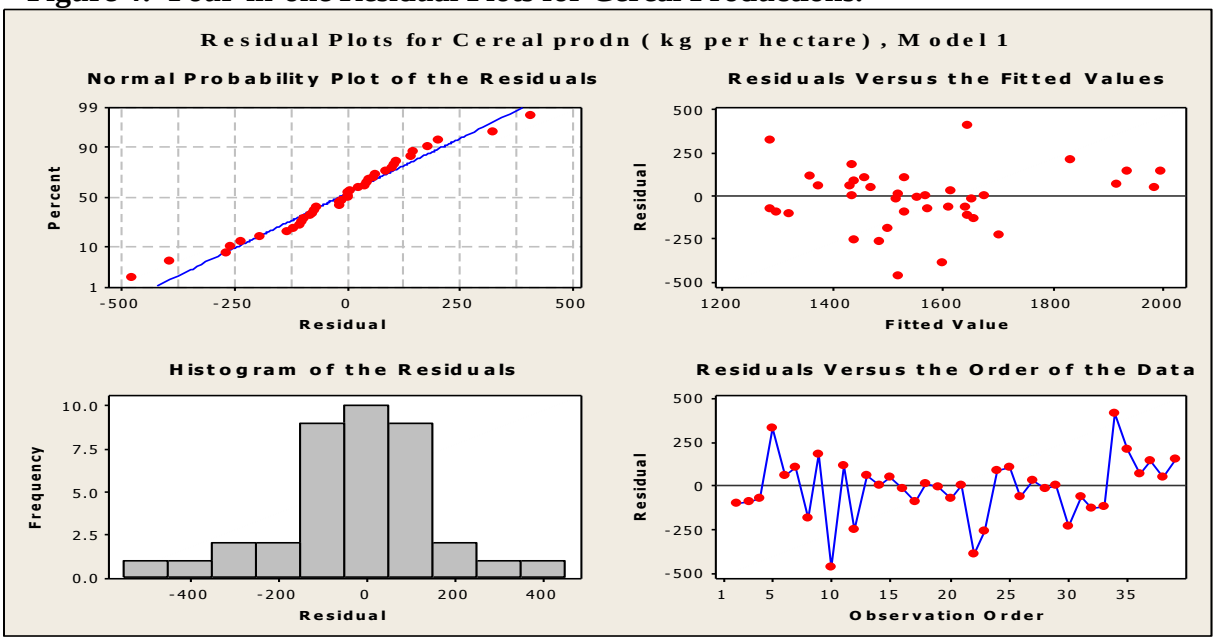
To evaluate the forecasting performance of this model, we used Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-Squared values as shown in table 6. Based on all these criteria, clearly, we can see that ARIMA (1,1,2) model has smaller values and out performs the ARIMA (1,1,1) and ARIMA (1,1,0) respectively.

Table 6: Final Estimates of Parameters for Model 1

Type	Coef	SE Coef	T	P
AR 1	0.7353	0.1359	5.41	<0.0001
MA 1	1.3817	0.0805	17.17	<0.0001
MA 2	-0.4100	0.1218	-3.37	0.002
Constant	4.981	1.440	3.46	<0.0001

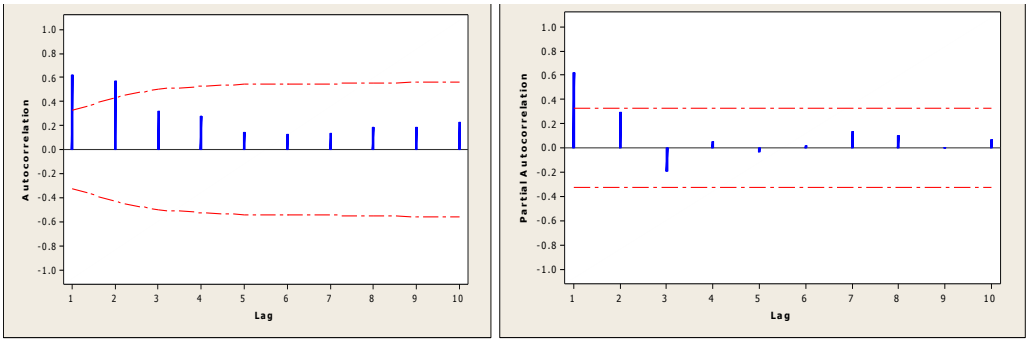
The predicted average annual cereal yields can therefore be determined by ARIMA (1,1,2) confirmed by the significance of all its parameters.

Figure 4: Four-in-one Residual Plots for Cereal Productions.



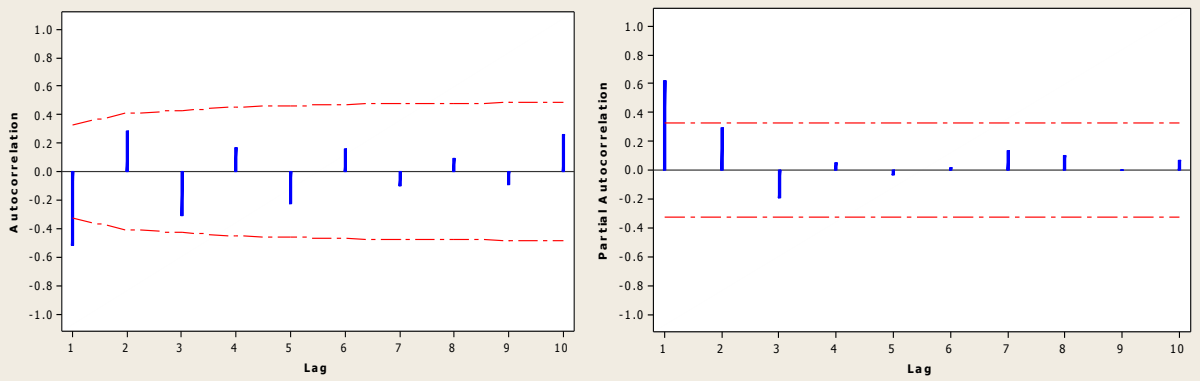
The four –in-one residual plots in Figure 4 above confirms the fact that all the parameters including the constant are significantly different from zero, the p-values of both the Moving Average and Autoregressive models are significantly smaller than 0.05. Considering randomness on the residuals. The null hypothesis is that, the residuals are white noise and hence since the p-values of Ljung-Box are greater than 0.05, we will not reject the null hypothesis, it indicates that the residuals behave like white noise and as such our fitted model seems appropriate.

Figure 5: ACF and PACF Plots on Cereal Production for Model 1



The Minitab output in Figure 5 above confirms that an ARIMA (1,1,2) is the appropriate model for this series. We do acknowledge that ACF dies positively very slowly with increases in the lags while PACF cuts off after lag 1. The displayed Ljung-Box test also supports the variations and the non-stationarity realized due to a boost in Cereal productions from modern agriculture. The sample autocorrelation function (ACF) of the residuals further shows that the residuals are approximately correlated, and therefore a presence of a model in the process. Autocorrelations at low lags are very high, and decline slowly as the lag increases for a trended series. This deterministic trend can be removed by using a "detrending" approach of taking first or second differences in order to transform this non-stationarity in cereal production time series data. However, this approach requires an assumption that the trend lasts "forever" and ignores the correlation among time lags since the deterministic trend models are based on the least squares model fit.

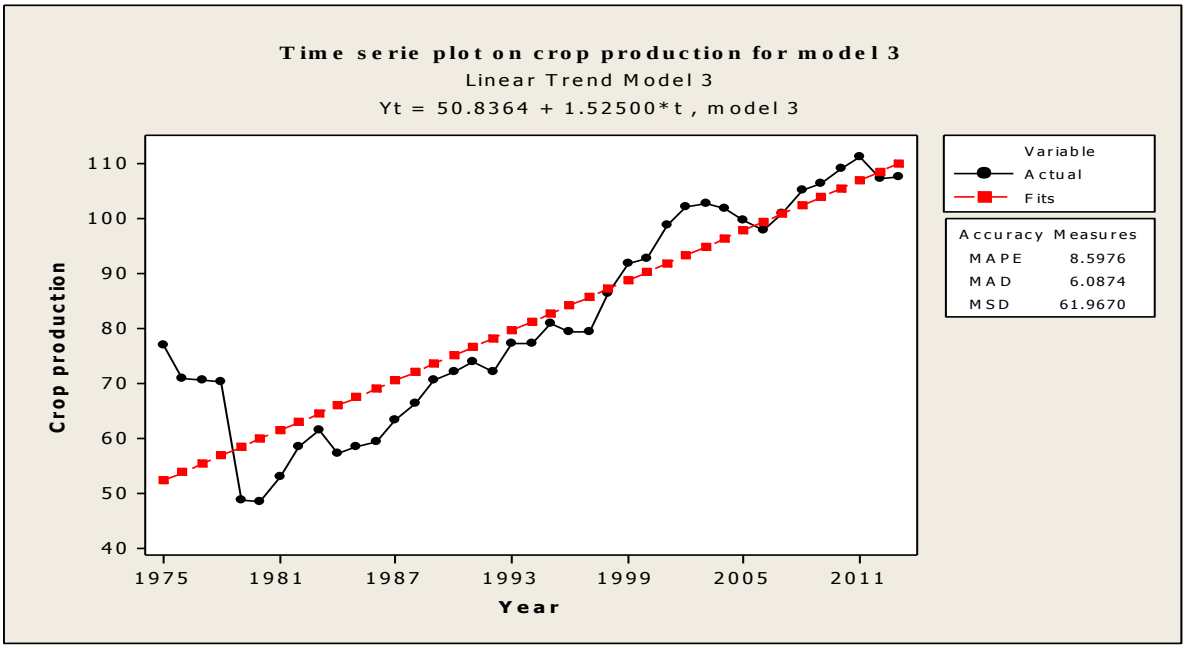
Figure 6: ACF and PACF Plots on Cereal Production after Differencing for Model 2



Differencing the series once cause the ACF to change and cut off negatively after lag 1 behaving like a random walk while PACF remained unchanged after differencing once which is an indication that the ARIMA (1,1,2) model is appropriate for the series and the parameters should be significant.

Crop production is the second major agricultural output in Uganda with a steady increase in its productivities since the late 80s with the applications of modern agricultural practices. The major crop has always been coffee that earned Uganda major foreign currencies from its exports as well as boosting the efforts to reduce or eradicate multidimensional poverty in Uganda. It is apparent that the increasing linear trend shown on Figure 7 also witnessed some yearly seasonal fluctuations as a result of falls in prices of coffee in the international market.

Figure 7: Trend Analysis for Crop Production (2004-2006 = 100).



5. Agricultural Share to GDP

Time series data is used to analyze the share of agriculture to the GDP of Ugandan economy with its potential roles to multidimensional poverty reductions. Regression analysis (OLS) is performed to get desired results and useful statistical information from the analysis or study. To test this hypothesis empirically, the role of agriculture on the GDP and its correlation to poverty, the model can be specified as;

$Y=\beta_0 +\beta_1 Agri + \beta_2 Ind +\beta_3 Trade +\beta_4 Manuf + Є$

Where,

Y = Gross Domestic Product of Uganda (billion dollars)

Agri = Agricultural

Ind = Industry

Trade = Trade

Manuf = Manufacturing

Table 7: GDP verses Agriculture, Trade, Industry and Manufacturing for Model 4

Predictor	Coef	SE Coef	T	P	VIF
Constant	155.0	310.4	0.50	0.621	
Agriculture	-0.4165	0.3531	-1.18	0.246	18.3
Trade	0.7237	0.2247	3.22	0.003	3.3
Industry	-1.6445	0.7945	-2.07	0.046	15.2
Manufacturing	5.1378	0.9982	5.15	<0.0001	2.4

MSE = 1450.71 **R-Sq** = 78.2% **R-Sq(adj)** = 75.7%

Table 8: Analysis of Variance for Model 4

Source	DF	SS	MS	F	P
Regression	4	257322994	64330749	30.57	<0.0001
Residual Error	34	71554847	2104554		
Total	38	328877841			

The value of R-square is a statistical measure of how much the independent variable explains the variations in the dependent variable. This indicates that approximately 78% variation in Gross domestic product (GDP) of Uganda is explained by the explanatory variables (Agriculture, Industry, Trade and Manufacturing).

The significance of F. value shows that in over all, models is good fitted. As per expectation, the GDP of Uganda is positive but affected by declining shares from agriculture sector which play an important role. GDP is incorporated to take into custody the factors associated with the intensity of economic growth.

6. Conclusions

The prospects for agricultural development in Uganda are bright. They are better now than several years ago in Uganda compared to Asia and Latin America during the period when those regions were going through agricultural revolutions. Much empirical evidence in Uganda is for poverty reduction through increases in modern agricultural productivity and one can conclude that:

Modernization of agriculture will be an ultimate solution to multidimensional poverty reduction in Uganda directly by increasing real household incomes, indirectly through employment generation and food price effects. Evidences support the theories that when farm incomes and the real wage rate increase and the rural non-farm economy grow, real household incomes increase and the percentage of the population living below international poverty lines decreases.

7.References

Alkire, S., and M. Santos. (2010). "Acute Multidimensional Poverty: A New Index for Developing Countries." Human Development Research Paper 11. UNDP–HDRO, New York.

Appleton, S. (2001). 'Changes in Poverty and Inequality', Chapter 4 in P. Collier and R. Reinnikka (eds.) Uganda's Recovery: The Role of Farms, Firms and Government pp.83-121, World Bank: Washington DC.

Eastwood, R., M. Lipton and A. Newell (2010). Farm Size. In Pingali, P. and E. Evenson (eds.): Handbook of Agricultural Economics. North Holland: Elsevier.

Flygare, S. (2006). The Cooperative Challenge: Farmer Cooperation and the Politics of Agricultural Modernization in 21st Century Uganda. Uppsala: Uppsala Studies in Economic History.

Fuglie, K. (2009): Agricultural Productivity in Sub-Saharan Africa. USDA, Economic Research Service. Paper presented at the Cornell University Symposium on the Food and Financial Crises and their Impact on the Achievement of the Millennium Development Goals, New York.

Nabudere, D. W. (1997). 'Beyond Modernization and Development, or Why the Poor Reject Development', Geografiska Annaler. Series B, Human Geography, Vol. 79.

Republic of Uganda (2013). 'Uganda National Panel Survey 2011/2012, UBOS Publication.

Ruecker, G., S. Park, H. Ssali and J. Pender (2003). Strategic Targeting of Development Policies to a Complex Region: A GIS-Based Stratification Applied to Uganda. ZEF Discussion Paper 69, Bonn: Center for Development Research, University of Bonn.

Uganda Bureau of Statistics (UBOS) and Macro International Inc. (2007). Uganda Demographic and Health Survey 2006. Calverton, Maryland, USA: UBOS and Macro International, Inc.

USAID (2007). Uganda: Formative Evaluation of the Agricultural Productivity Enhancement Program (APEP), Final Report. Washington, D.C

Wiggins, S. and Leturque, H. (2010). Helping Africa to Feed Itself: Promoting Agriculture to Reduce Poverty and Hunger. London, ODI, World Bank.

World Bank (2009). World Development Indicators. Washington DC: World Bank

------- (2012). 'Uganda: Promoting Inclusive Growth, Synthesis Report.
Washington DC.

جامعة القاهرة
كلية الدراسات العليا للبحوث الإحصائية

المؤتمر السنوى الرابع والخمسون للإحصاء
وعلوم الحاسب وبحوث العمليات

إحصاء تطبيقى

9-11 ديسمبر 2019