

Cairo University
Institute of Statistical Studies and Research

**The 52nd Annual Conference on Statistics, Computer
Sciences and Operations Research**

Applied Statistics

25-27 Dec. 2017

Index

APPLIED STATISTICS

1	On Quantile Regression for Dynamic Panel Data Models الحسينى عبد البر ، أحمد الشيخ ، حسام الدين محمود أحمد	1-15
2	A note on Markov Models and Chronic Diseases الحسينى عبد البر ، أيمن سعد الدين	16-33
3	Subset Selection in Multiple Linear Regression (Review) أحمد الشيخ ، أسماء وهبه	34-48
4	Estimation of Unbalanced Panel Data الحسينى عبد البر ، أحمد الشيخ ، آيه محمود أحمد	49-56
5	Extracting the Wealth of Knowledge in Composite Index Data أسماء محمد سيد	57-79
6	Review: Conditional and Unconditional AR(1) Model سيد مشعال ، أحمد أمين ، محمد خليفه ، هادية فريد	80-87
7	Estimation of Ranked Set Sampling Methods : Comparative Study نصر إبراهيم رشوان ، نوران مدحت الموان	88-103

On Quantile Regression for Dynamic Panel Data Models

Elhoussainy A. R.¹

Ahmed A.E.²

Hossameldin M. A.³

Abstract

This paper studies a quantile regression dynamic panel model with fixed effects. Panel data fixed effects estimators are typically biased in the presence of lagged dependent variables as regressors. To reduce the dynamic bias, we review the suggestion of the use of the instrumental variables quantile regression method of (Chernozhukov & Hansen, 2008) along with lagged regressors as instruments.

Keywords: Quantile regression, Dynamic panel, fixed effects, Instrumental variables

1. Overview on Quantile Regression (QR)

Quantile regression is a statistical technique able to detect more effects than conventional procedures: it does not restrict attention to the conditional mean and therefore it permits to approximate the whole conditional distribution of a response variable. It mainly aims at estimating conditional quantile of a certain variable of study given a set of covariates. Therefore, in the presence of patterned dataset, skewed distributed dataset or outliers, it would be possible to

figure out the effect of set of independent variables on a certain conditional quantile of a certain outcome with variant estimates (Chetverikov, et al., 2016)

A breakthrough in regression analysis is the quantile regression approach proposed by (Koenker & Bassett, 1978). This approach permits estimating various quantile functions of a conditional distribution, among them the median (0.5th quantile) function is a special case. Each quantile regression characterizes a particular (center or tail) point of the conditional distribution; putting different quantile regressions together thus provides a more complete description of the underlying conditional distribution. This analysis is particularly useful when the conditional distribution is heterogeneous and does not have a “standard” shape, such as an asymmetric, fat-tailed. (Powell, 1986) studied quantile regressions for censored data, and (Koenker , 2000) studies Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics

Quantile regression has several unique features than other regressions. It can be utilized for characterizing the whole conditional distribution of a dependent variable. The estimation is performed by Linear Programming (LP) which can be considered to be relatively easier than other conventional estimation methods. It provides a robust measure of location for non-symmetrically distributed dataset. Quantile regression is relatively able to grant a more efficient estimator than OLS method when error term normality assumption is violated. *L*-estimators (robust estimator based on linear combination of order statistic to be resistant to outliers) which depend on a linear combination of

quantile estimators are relatively more efficient than Least squares estimators (Portnoy & Lin, 2010)

Quantile regression aims at estimating conditional quantile of a certain variable of study given a set of regressors. Therefore, in the presence of patterned dataset, skewed distributed dataset or outliers, it would be possible to figure out the effect of set of independent variables on a certain conditional quantile of a certain outcome. That means for a certain varying outcome variable, several regression lines can be estimated to be the presentation of every quantile of a dependent variable (Chernozhukov & Fernández, 2011).

It is able to identify the behavioral effect of different regressors on a certain variable of study. That would help recognize if there is any change of their effects at several quantiles or stable effect over the whole distribution (Chetverikov, et al., 2016)

Conceding that, $\theta \in [0, 1]$ that would generate several estimated parameters according to variant values of θ . In other words, at the θ^{th} quantile, the coefficient ($\hat{\beta}_{\theta i}$) can be interpreted as the amount of change in the variable of study when a certain regressor changes by a one unit holding the rest of regressors fixed. That would allow researchers to investigate the effect of a set of covariates on the variable of interest at several quantiles. Studying the effects on a dependent variable at several levels of quantile, can indicate whether the effect is relatively the same cross the quantiles or it is variant which denotes the different influences on variable of interest at the low quantiles than at the high

ones. Detecting these effects cross quantiles is a considerably benefited from utilizing quantile regression.

An estimate for θ^{th} quantile of y can be obtained by minimizing the sample analoug of population objective function as follows

$$\hat{\beta}_{\theta} = \min_{\beta} \frac{1}{n} \sum_{i=1}^N \rho_{\theta}(Y_i - X_i' \beta) \quad (1)$$

Where $i = 1, 2, \dots, n$, $\{y_1, \dots, y_n\}$ is a random sample of response variable Y , x_i is the covariate vectore corresponding to the i^{th} observation, ρ_{θ} is an indicator function $I(\cdot)$ defined to be θ for $i: y_i \geq x_i' \beta$ and $(1 - \theta)$ for $i: y_i < x_i' \beta$. $\hat{\beta}_{\theta}$ is the estimated parameters at a particular level of θ . The minimization analog in (1) can be presented in an alternative form to be as following:

$$\hat{\beta}_{\theta} = \min_{\beta} \left[\sum_{i: y_i \geq x_i' \beta}^N \theta |y_i - x_i' \beta| + \sum_{i: y_i < x_i' \beta}^N (1 - \theta) |y_i - x_i' \beta| \right] \quad (2)$$

Where $i = 1, 2, \dots, N$

In view of (2), the θ^{th} quantile regression estimator of β can be obtained by minimizing its sample counterpart, i.e., the average of asymmetrically weighted absolute errors with weight θ on positive errors ($u_i: y_i > x_i' \beta$ and weight $(1 - \theta)$ on negative errors $u_i: y_i < x_i' \beta$. Accordingly, LAD is obtained at $\theta = 0.5$ where the median regression locates. The argument in (2) can also be defined as

$$\hat{\beta}_{\theta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \rho_{\theta}(y_i - x'_i \beta) \quad (3)$$

where Median estimator (L_1) can be obtained in a certain special case where $\theta = 0.5$ which refers to the median estimator.

The function in (2) is minimized several linear programming methods, which can guarantee to yield a solution in a finite number of iterations. Quantile regression masters the advantage over OLS. OLS can be inefficient if the errors are highly non-normal, QR is more robust to non-normal errors and outliers. QR also provides a richer characterization of the data, permitting considering the impact of a covariate on the entire distribution of y , not merely its conditional mean. This non-differentiable function is minimized via the simplex method, which is guaranteed to yield a solution in a finite number of iterations **Invalid source specified.** and (Chetverikov, et al., 2016)

Since the fundamental goal of quantile regression is to estimate the conditional quantile of a response variable given a set of covariates at certain quantile of θ , therefore, the model in (1) can be presented in the form of θ^{th} conditional quantile of y on a $K \times 1$ vector of covariates $x = (1, x_1, \dots, x_K)$

$$Q_{\theta}(y|x) = x' \beta_{\theta} \quad (4)$$

That would permit the conditional quantile term in (3) to be re-written as a conventional econometric model as following:

$$y = x' \beta_{\theta} + u_{\theta} \quad (5)$$

This model is considered to follow the traditional OLS assumptions. and the population vector of parameters can be presented as following:

$$\hat{\beta}_{\theta} = \min_{\beta} E[\rho_{\theta}(u_{\theta}|x)]$$

Or alternatively,

$$\hat{\beta}_{\theta} = \min_{\beta} E[\rho_{\theta}(y - x' \beta|x)] \quad (6)$$

Within the frame of sample, the minimization argument at θ^{th} quantile for conditional quantile can be defined in the same sense, given $(y_i, x_i), i = 1, 2, \dots, n$ which is a i.i.d sample drawn from a certain population, the vector of parameters can be defined as following:

$$\hat{\beta}_{\theta} = \min_{\beta} \frac{1}{n} \sum_{i=1}^n \rho_{\theta}(u_{\theta}) \quad (7)$$

Or it can be defined within sample frame in an alternative way as following:

$$\hat{\beta}_{\theta} = \min_{\beta} \frac{1}{n} \left[\sum_{i: y_i \geq x_i' \beta}^n \theta |y_i - x_i' \beta| + \sum_{i: y_i < x_i' \beta}^n (1 - \theta) |y_i - x_i' \beta| \right]$$

2. Dynamic Panel Data Model with Fixed Effects (DPDFE)

Dynamic panel data (DPD) has increasingly become a main pillar in several real life applications, especially in Economics field. It provides the privilege of lagged dependent variable inclusion which permits more explanation to time invariant nature of several variables of studies. The conventional

dynamic panel data model would involve a cross section specific effect for a particular cross section, as a result of that, this time-invariant term would provide the differential for a specific cross section.

(Galvao, 2011) considered the following DPD-FE model of cross section units of $i = 1, 2, \dots, N$ and time periods of $t = 1, 2, \dots, T$.

$$Y_{it} = J(Y_{it-1}, X'_{it}, \mu_i, \varepsilon_{it})$$

$$Y_{it} = \delta Y_{it-1} + X'_{it} \beta + \mu_i + \varepsilon_{it} \quad (9)$$

Where Y_{it} is the variable of study or response variable, δ is the parameter of lagged dependent variable, and for stationarity purposes, $|\delta| < 1$ should be satisfied. X'_{it} is a P – dimensional row vector of continuous time-variant exogenous variables, β is a P - dimensional column vector of parameters for exogenous variables. μ_i represents cross sectional individual specific-effect and ε_{it} is innovation classical error term and the endogenous lagged dependent variable Y_{it-1} . The model in (9) is considered –strictly - to be DPD fixed effect model if mainly the correlation between individual effect μ_i and regressors is present. The model is presented with taking the following assumptions into consideration:

Assumption (1): across the panels, the error terms are distributed as $\varepsilon_{it} \sim iid(0, \sigma_\varepsilon^2)$

Assumption (2): the error term observations are not correlated with the explanatory variables, so exogeneity is present and $E(x_{it}, \varepsilon_{it}) = 0$.

Assumption (3): the exogenous variables are correlated with the individual effect, so $E(x_{it}, \mu_i) \neq 0$.

Assumption (4): the error terms are correlated with the lagged endogenous variable $E(y_{it-1}, \varepsilon_{it}) = 0$

Assumption (5): across the panels the initial value of Y_{i1} is abtained by

$Y_{i1} = \frac{\mu_i}{1-\delta} + G_{i1}$, where $i = 1, 2, \dots, N$, $G_{i1} = \sum_{m=0}^{\infty} \delta^m \varepsilon_{it-1}$ and it is in no correlation with individual specific effect μ_i .

Least Square Dummy variable (LSDV), within and between groups estimators are considered to be the main estimation methods for DPD-FE models. To eliminate the effect of individual effect, the first difference is the transformation technique that is used in most of early literature. (Nickell, 1981) showed that wiping this term can be performed by taking difference approach follows:

$$\begin{aligned} Y_{it} - Y_{it-1} &= \delta(Y_{it-1} - Y_{it-2}) + (X_{it} - X_{it-1})\beta + (\mu_i - \mu_i) + (\varepsilon_{it} - \varepsilon_{it-1}) \\ \Delta Y_{it} &= \delta \Delta Y_{it-1} + \beta \Delta X_{it} + (\mu_i - \mu_i) + \Delta \varepsilon_{it} \end{aligned} \quad (10)$$

Since the individual effect term is time invariant, therefore it will be removed. As long as the time-invariant individual effect μ_i is a component in the functions of variable of study Y_{it} and Y_{it-1} , therefore, Y_{it-1} is correlated with the idiosyncratic error term U_{it} . As a result of that, LSDV would obtain inconsistent and certainly biased estimators even if error term observations are not in serial correlation. (Belloni & Chernozhukov, 2011) have also showed that, despite the

usage of LSDV (FE method), the existence of Y_{it-1} will maintain the correlation with the error term observations which would produce a bias of $O(1/T)$.

Handling this bias has been extensively studied by (Anderson & Hsiao, 1981). They have provided two main approaches to treat the arising problem of endogeneity from the differenced term of ΔY_{it-1} ; they depend on replacing it by a term which is more correlated with the variable of study and in no correlation with the error term. They proposed the essential approaches of considering lagged difference (ΔY_{it-2}) and lagged value (Y_{it-2}) of Y_{it-1} as these instruments are in no correlation with $\Delta \varepsilon_{it}$. In real life, it is quite hard to find an instrument for the lagged difference value of dependent variable. (Kiviet, 1995) showed that DPD-FE models that utilize instrumental variable estimation would suffer from weak finite sample efficiency and bias of $O(N^{-1} T^{-3/2})$.

3. Quantile Regression for DPD With Fixed Effects (QRDPDFE)

The estimation of quantile regression in dynamic panel data with fixed effects (QRPFE) has a growing literature over the last decade. The presence of individual effects is based on the assumption that there is a correlation between cross section-specific effect and regressors. Individual effects can also manage the heterogeneity across the panels. In addition to that, the inclusion of instrumental variable(s) due to the existence of lagged regressor has become a new approach. Recall the model in (9), it can be put in concise matrix form as follows:

$$Y = Z\mu + \delta Y_{-1} + X\beta + \varepsilon$$

Z is an incidence matrix which represents the parameters of separate individuals (N) in the sample of interest, and is obtained by $I_N \otimes l_T$ is a $T \times 1$ is a ones-vector. $\mu = (\mu_1, \dots, \mu_N)'$ is the $N \times 1$ individual specific effects vector which represents the intercepts.

The model in (9) can be presented in the frame of quantile regression. Considering i th individual and t th time period observations and at a particular level of quantile (θ) the model can be as follows:

$$Q_{y_{it}}(\theta | y_{it-1}, x_{it}) = \delta(\theta) Y_{it-1} + X'_{it} \beta(\theta) + \mu_i \quad (11)$$

Both y_{it-1} and x_{it} are the only regressors that rely on quantile level of (θ). And as long as this is a fixed effect model, therefore it permits the correlation between individual effect and covariates. Some of the variability in the response variable y_{it} is interpreted by the inclusion of μ 's. Involving them would

allow more controlling for unobserved heterogeneity that cannot be sufficiently interpreted by other regressors.

μ 's denote the time-invariant covariate and their parameters represent pure location shift effect on the variable of study y_{it} . Panel data would acquire its important nature from the excess number of cross sections over the length of given modestly smaller length of periods. In order to that, it would be considerably difficult, unrealistic and meaningless to estimate θ th distributional shift $\mu_{i\theta}$ for each individual, as line might not pass adequately due to the relatively small time series length. Hence, to avoid this problem, it has been suggested by (Koenker, 2000) to assume that the estimation of individual effect is independent from the level of quantile (θ). This assumption can be performed by estimating the model in (11) for several quantiles simultaneously. Based on (Koenker, 2004), the model in (11) can be presented as follows:

$$\min_{\mu, \delta, \beta} \sum_{s=1}^S \sum_{i=1}^N \sum_{t=1}^T p_s \rho_{\theta} (Y_{it} - X'_{it} \beta(\theta) - \delta(\theta) Y_{it-1} - \mu_i) \quad (12)$$

Where $\rho_{\theta}(u) := u(\theta - I(u < 0))$ as defined by Koenker and Basset (1978), the weights p_s control the potential relative effect of a given S quantiles $(\theta_1, \theta_2, \dots, \theta_S)$, the selection of the weights and the related quantiles are based on Weighted L-statistic according to (Mosteller & Tukey, 2002).

4. Quantile Regression for Dynamic Panel Data with Fixed Effects - Instrumental Variables Approach (QRDPDFEIV)

In dynamic panel data, the presence of lagged value of dependent variable will lead to biased estimators would generate biased estimators due to the correlation between error term and lagged dependent variable. The biased estimators would result from OLS and QR techniques.

Within the conceptual frame of linear regression estimation, Anderson and (Hsiao, 1986) and (Arellano & Bond, 1991), provided several IV based-frameworks that can help provide consistent estimators in DPD models. Based on that objective, they provided the perspective of utilizing lagged or lagged differences of the endogenous variables of interest. The exploitation of these instrument(s) extract its importance from being correlated with their endogenous variables but not with the innovation term. As a result of that, the bias can be minimized and obtain more consistent estimators.

The involvement of a particular valid instrument (m_{it}) in the model as a part of regressors is the traditional method of estimating a specific parameter(s). By considering the minimization argument in (12), this frame can be presented as follows:

$$\min_{\mu, \delta, \beta, \lambda} \sum_{s=1}^S \sum_{i=1}^N \sum_{t=1}^T p_s \rho_{\theta} (Y_{it} - X'_{it} \beta(\theta) - \delta(\theta) Y_{it-1} - \lambda(\theta) m_{it} - \mu_i) \quad (13)$$

Where, Y_{it-1} is a $\dim(\delta)$ -vector of endogenous variable, μ_i is the FE term, X'_{it} is a $\dim(\beta)$ -vector of exogenous covariates, m_{it} is a $\dim(\lambda)$ -vector of instrumental variables. And it should be taken into consideration that $\dim(\lambda) \geq \dim(\delta)$. The involvement of Y_{it-1} would generate bias, however, it can be trimmed by involving IV (m_{it}), that can influence the variable of interest Y_{it} , nevertheless it is independent from ε_{it} . In Case of acquiring valid instruments, therefore they will assist in estimating δ and β in a better way.

The lagged or lagged differences values for two time periods or more of (Y_{it}) and (X_{it}) can be probably used to as instruments (m_{it}) to estimate δ and/or β .

In real life, one of two main basic techniques is followed. The first technique is based on running (m_{it}) – which is/are selected based on the theory- along with the other covariates to gain the corresponding effects on Y_{it} .

The second technique depends on estimating (m_{it}) by using the predicted values from OLS projection of lagged Y on m and X , which are considered to be independent from the error term as follows:

$$\tilde{y}_{i-} = g(m_{it}, x_{it}, v_{it}) \quad (14)$$

And then to insert it in the following model:

$$Y_{it} = \delta \tilde{Y}_{i-} + X'_{it} \beta + \mu_i + \varepsilon_{it} \quad (15)$$

The two approaches are useful to minimize the bias effect. And they can be stemmed from the economic theory (inside the model) or can be estimated by one of the mentioned techniques.

5. Conclusion

We presented the evolvement of applying the quantile regression in dynamic panel data. In addition to that we reviewed the recent developments in handling the bias problem in DPDFE models by utilizing IV approach. This approach is intended to minimize the presence of bias in DPDFEIV models.

6. References

- Alizadeh , M., Bagheri, S., & Nadarajah, S. (2015). Efficient Estimation of the PDF and the CDF of the Exponentiated Gumbel Distribution. *Communication in Statistics - Theory and Methods*, 45(1), 339-361.
- Anderson, T., & Hsiao, C. (1981). Estimation of dynamic models with error components. *JASA*, 589-606.
- Arellano, M., & Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to. *The Review of Economic Studies*, 58(2), 277-297.
- Bassett, G., Koenker, R., & Kordas, G. (2004). Pessimistic Portfolio Allocation and Choquet Expected Utility. *Journal of Financial Econometrics*, 2, 477–492.
- Belloni, A., & Chernozhukov, V. (2011). ℓ_1 -penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1), 82-130.

- Chernozhukov, V., & Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1), 379-398.
- Chetverikov, D., Larsen, B., & Palmer, C. (2016). IV Quantile Regression for Group-Level Treatments, With an Application to the Distributional Effects of Trade. *Journal of Econometric Society*, 84(2), 809-833.
- Galvao, A. (2011). Quantile regression for Dynamic Panel Data with Fixed Effects. *Journal of Econometrics*, 164, 142-157.
- Hsiao, C. (1986). *Analysis of Panel Data*. New York: Cambridge University Press.
- Kiviet, J. F. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics*, 68(1), 53-78.
- Koenker, R. (2000). Galton, Edgeworth, Frisch, and prospects for quantile regression in Econometrics. *Journal of Econometrics*, 95, 347-374.
- Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1), 74-89.
- Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1), 33-50.
- Mosteller, F., & Tukey, J. (2002). Data Analysis and Regression. *The Annals of Statistics*, 30(6), 1535–1575.
- Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*, 49(6), 1417-1426.
- Portnoy, S., & Lin, G. (2010). Asymptotics for Censored Regression Quantiles. *Journal of Nonparametric Statistics*, 22(1), 115-130.
- Powell, J. (1986). Censored regression quantiles. *Journal of Econometrics*, 32, 143-155.

A note on Markov Models and Chronic Diseases

El-Houssainy A. Rady*

Ayman S. Anwar**

Abstract

The paper is concerned with introducing the concept of Markov model concentrating on its relationship with chronic diseases. An introduction of Markov models is given in section one, the Hidden Markov model is explained in section 2. Section 3 is concerned with kidney disease and its stages, Support Vector Machine classification technique described in section 4.

Keywords: Markov Model, Hidden Markov Model, Chronic Kidney Disease, Data Mining techniques, Support vector machine.

Introduction

Chronic diseases are often described by stages of severity. Clinical decisions about what to do are influenced by the stage, whether a patient is progressing, and the rate of progression.

The stage of disease and rate of progression or regression are important to deciding whether to treat, how to treat, and how often to monitor a patient. Moreover, knowledge about transition rates between stages helps patients understand what to expect and policymakers what to plan.

One approach for analyzing disease stage data is hidden Markov models (HMMs) (MacDonald and Zucchini 1997, 2009). Unlike ordinary Markov models, HMMs account for the fact that sometimes the observed disease stages are different from the underlying disease stages as a result of measurement error.

Institute of Statistical Studies & Research

* Professor of Applied Statistic, Institute of Statistical Studies & Research.

** Corporate MIS Manager, Magrabi Hospitals & Centers.

Recently, researchers have used continuous-time HMMs to analyze data in a variety of clinical areas, such as hepatocellular cancer (Kay 1986), HIV progression (Satten and Longini 1996), and aortic aneurysms (Jackson 2003). However, a continuous-time model is computationally costly, and may be infeasible if the sample size is large, which is typically the case with electronic health records (EHR) data. Further, for many studies there would be no benefit to having finer information about the timing of a measurement than the calendar date. Discrete-time HMMs are a useful alternative, and have been developed and applied to a variety of health problems (Shirley et al. 2010; Rabiner 1986; Jackson and Sharples 2002; Scott 1999; Scott 2002; Scott et al. 2005; Gentleman et al. 1994; Bureau et al. 2000).

While discrete-time HMMs have many desirable features, the estimation of transition rates typically requires large observational studies with long follow-up times as transitioning usually occurs over years. The resources required for such studies are often costly and time prohibitive. Use of longitudinal EHRs data from large primary care practices offers an alternative means of assembling longitudinal health experience of a population. Such data have the advantage of having both sufficient follow-up time and sample size to reliably and accurately estimate these rare transition rates.

1. Markov Models

Markov model is a stochastic model used to model randomly changing systems where it is assumed that future states depend only on the present state and not on the sequence of events that preceded it (that is, it assumes the Markov property). Generally, this assumption enables reasoning and computation with the model that would otherwise be intractable.

Markov chain

It is a stochastic model described as follows: we have a chain of a set of states, $S = \{s_1, s_2, \dots, s_n\}$. the process starts in one of these states and moves successively from one state to another. Each move is called a step. If the chain is currently in state s_m then its moves to state s_n at the next step with a probability p_{mn} and this probability doesn't depend on which state was the chain in before the current state.

The Markov property suggests that the distribution for this variable depends only on the distribution of the previous state.

Formally, A Markov chain is a sequence of random variables $W_1, W_2, W_3, \dots$ with the Markov property, namely that, given the present state, the future and past states are independent.

The components of Markov Chain

- State space denoted by S which is a set of all possible state of the chain
- Transition probabilities denoted by P which is a set of all possible transition probabilities to move from each state to another state or stay on the same state.

Example-1:

Suppose that a diabetic patient has three stages of disease severity: Normal Stage, Medium Severity Stage and Critical Severity Stage. Let's assume for the moment that the severity of disease doesn't change during the day.

Disease severity prediction is trying to guess what the stage of disease will be next day based on a history of observations of disease severity stages., we can express that by the following probabilities:

$$P(s_n | s_{n-1}, s_{n-2}, s_{n-3}, \dots, s_1) \quad (1)$$

Using formula (1), we can give probabilities of diabetic severity stages for next day using n days of history., but the problem here that for large n more statistics must be collected, if we suppose that $n=10$, then we have to collect statistics for $3^{10}=59,049$ past histories, so we will simplify it by using Markov Assumption:

$$P(s_n | s_{n-1}, s_{n-2}, s_{n-3}, \dots, s_1) \approx P(s_n | s_{n-1}) \quad (2)$$

Also, we can express the joint probability using the Markov Assumption as follow:

$$P(s_1, s_2, s_3, \dots, s_n) = \prod_{i=1}^n P(s_i | s_{i-1}) \quad (3)$$

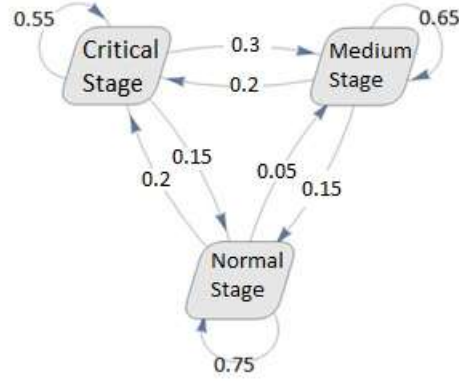
Let's arbitrarily pick some numbers for probabilities of next day diabetic severity stage based on today severity stage (Transition Probabilities P), expressed in table (1)

Table 1. Transition Probabilities for diabetic severity stages

		Next day Diabetic Severity Stage		
		Normal Stage	Medium Severity Stage	Critical Severity Stage
Today's Diabetic	Normal Stage	0.75	0.05	0.2
	Medium Severity Stage	0.15	0.65	0.2
	Critical Severity Stage	0.15	0.3	0.55

$$P = \begin{pmatrix} 0.75 & 0.05 & 0.2 \\ 0.15 & 0.65 & 0.2 \\ 0.15 & 0.3 & 0.55 \end{pmatrix}$$

Figure -1. Diabetic severity stage transition diagram



So, given that today diabetic patient in a normal stage it's easy compute the Probability that tomorrow will be also in normal stage and the day after tomorrow will be in a critical stage as the following: -

$$\begin{aligned}
 P(s_3=\text{critical}, s_2=\text{normal} \mid s_1=\text{normal}) &= P(s_3=\text{critical} \mid w_2=\text{normal}, w_1=\text{normal}) * \\
 &\quad P(s_2=\text{normal} \mid s_1=\text{normal}) \\
 &= P(s_3=\text{critical} \mid s_2=\text{normal}) * \\
 &\quad P(s_2=\text{normal} \mid s_1=\text{normal}) \\
 &= 0.2 * 0.75 \\
 &= 0.15
 \end{aligned}$$

Also, the probability that the diabetic patient on the day after tomorrow will be in a medium stage given that today severity stage is critical can be computed as follow: -

$$\begin{aligned}
 P(s_3=\text{medium} \mid s_1=\text{critical}) &= P(s_3=\text{medium}, s_2=\text{critical} \mid s_1=\text{critical}) + \\
 &\quad P(s_3=\text{medium}, s_2=\text{medium} \mid s_1=\text{critical}) + \\
 &\quad P(s_3=\text{medium}, s_2=\text{normal} \mid s_1=\text{critical}) +
 \end{aligned}$$

$$\begin{aligned}
 &= P(s_3 = \text{medium} \mid s_2 = \text{critical}) P(s_2 = \text{critical} \mid s_1 = \\
 &\text{critical}) + \\
 &P(s_3 = \text{medium} \mid s_2 = \text{medium}) P(s_2 = \text{medium} \mid \\
 &s_1 = \text{critical}) + \\
 &P(s_3 = \text{medium} \mid s_2 = \text{normal}) P(s_2 = \text{normal} \mid s_1 = \\
 &\text{critical}) \\
 &= 0.3 * 0.55 + 0.65 * 0.3 + 0.05 * 0.15 \\
 &= 0.165 + 0.195 + 0.0075 \\
 &= 0.3675
 \end{aligned}$$

It can be also computed directly by calculate $P^{(2)}$

$$\begin{aligned}
 P^{(2)} &= \begin{pmatrix} 0.75 & 0.05 & 0.2 \\ 0.15 & 0.65 & 0.2 \\ 0.15 & 0.3 & 0.55 \end{pmatrix} * \begin{pmatrix} 0.75 & 0.05 & 0.2 \\ 0.15 & 0.65 & 0.2 \\ 0.15 & 0.3 & 0.55 \end{pmatrix} \\
 &= \begin{pmatrix} 0.6 & 0.13 & 0.27 \\ 0.24 & 0.49 & 0.27 \\ 0.24 & 0.3675 & 0.3925 \end{pmatrix}
 \end{aligned}$$

$$P_{critical, medium}^{(2)} = 0.3675$$

Absorbing Markov Chains

A State s_i of a Markov chain is called absorbing if it is impossible to leave it (i.e., $p_{ii} = 1$)., A Markov Chain is *absorbing* if it has at least one absorbing state, and if from every state it's possible to go to an absorbing state within one or more step., also other non-absorbing states called *transient*.

Example – 2

Once the patient has clinical diabetic kidney disease, he can either remain in that health state or progress to higher stage kidney disease. Once the patient has contracted preliminary diabetic kidney disease, the model assumes he cannot be cured of it. Therefore, the patient either progresses to the clinical stage or remains in the preliminary stage. The following is a sample data of diabetic kidney disease transition probability matrix

Table 2. Transition Probabilities matrix for CKD stages

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Stage 1	0.93	0.06	0	0	0.01
Stage 2	0	0.85	0.12	0	0.03
Stage 3	0	0	0.75	0.15	0.10
Stage 4	0	0	0	0.65	0.35
Stage 5	0	0	0	0	1

$$P = \begin{pmatrix} 0.93 & 0.06 & 0 & 0 & 0.01 \\ 0 & 0.85 & 0.12 & 0 & 0.03 \\ 0 & 0 & 0.75 & 0.15 & 0.10 \\ 0 & 0 & 0 & 0.65 & 0.35 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

on that Case Stage 5 is *absorbing* State., Stage 2,...,Stage 4 are *transient*.

On such case (absorbing chains) it's interesting to determine the following: -

- What is the probability that the process will reach an absorbing state?
- What is the probability that the process will end up in a certain absorbing state?
- How long the process will take in average to be absorbed?

- How many times on average the process will be in each transient state?

Canonical Form

By reordering the transition probability matrix so that transient states come first and absorbing states on the next will get the following canonical form

$$P = \begin{matrix} \text{TRN} \\ \text{ABS} \end{matrix} \left(\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \end{array} \right)$$

Where: -

Q: is an $t \times t$ matrix, t is no of transient states.

R: is a none zero $t \times r$ matrix, r is no of absorbing states.

0: is an $r \times t$ zero matrix.

I: is an $r \times r$ identity matrix.

Fundamental Matrix

Definition (1) For an absorbing Markov chain **P**, the matrix $F = (I - Q)^{-1}$ is called the fundamental matrix for **P** and the entry n_{ij} of **F** gives the expected number of times that the process is in the transient state s_j if it is started in the transient state s_i .

$$P \text{ (Canonical form)} = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \left(\begin{array}{cccc|c} 0.93 & 0.06 & 0 & 0 & 0.01 \\ 0 & 0.85 & 0.12 & 0 & 0.03 \\ 0 & 0 & .75 & 0.15 & 0.10 \\ 0 & 0 & 0 & 0.65 & 0.35 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right)$$

$$F = \left[\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0.93 & 0.06 & 0 & 0 \\ 0 & 0.85 & 0.12 & 0 \\ 0 & 0 & 0.75 & 0.15 \\ 0 & 0 & 0 & 0.65 \end{pmatrix} \right]^{-1}$$

$$= \begin{pmatrix} 2 & 14.3 & 5.7 & 2.7 & 1.2 \\ 3 & 0 & 6.7 & 3.2 & 1.3 \\ 4 & 0 & 0 & 4 & 1.7 \\ 5 & 0 & 0 & 0 & 2.9 \end{pmatrix}$$

Time to Absorption

The Sum for all entries in the i^{th} row of F gives total expected number of times in any of transient states for a given starting state s_i which mean that, the expected time required before the chain being absorbed., Generally it can be computed as follow:

$$\mathbf{t} = \mathbf{F}\mathbf{c}$$

where: -

$\mathbf{t}$: column vector whose entries represent the expected time before the chain is absorbed given all possible starting states.

$\mathbf{F}$: Represent fundamental matrix.

$\mathbf{c}$: is a column vector all of whose rows are 1

$$\mathbf{t} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} \begin{pmatrix} 14.3 & 5.7 & 2.7 & 1.2 \\ 0 & 6.7 & 3.2 & 1.3 \\ 0 & 0 & 4 & 1.7 \\ 0 & 0 & 0 & 2.9 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 23.9 \\ 11.2 \\ 5.7 \\ 2.8 \end{pmatrix}$$

2. Hidden Markov Model

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. The mathematics behind the HMM was developed by L. E. Baum and coworkers 1966. It is closely related to an earlier work on the optimal nonlinear filtering problem by Ruslan L. Stratonovich 1960, who was the first to describe the forward-backward procedure.

In simple Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly.

Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics.

A hidden Markov model can be considered a generalization of a mixture model where the hidden variables, which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.

In its discrete form, a hidden Markov process can be visualized as a generalization of the Urn problem with replacement (where each item from the urn is returned to the original urn before the next step).

Example-3

Recall Example-1 and suppose that physician who follow-up diabetic patients is just come back from vacation, and he were asked about the stage of diabetic severity for examined patient. The only piece of evidence he has is the examined patient face status is it pale or not.

Let's assume the following probabilities: -

Table 3. Emission probabilities for seeing a pale face

	Probability of pale face
Normal Stage	0.1
Medium Stage	0.3
Critical Stage	0.8

Now we have to factor in the fact that the actual severity stage for the examined patient is *hidden* from him. We do that by using Bayes' Rule:

$$P(s_1, s_2, \dots, s_n | u_1, u_2, \dots, u_n) = \frac{P(u_1, u_2, \dots, u_n | s_1, s_2, \dots, s_n) P(s_1, s_2, \dots, s_n)}{P(u_1, u_2, \dots, u_n)}$$

Where u_i is true if the examined patient has a pale face on a day i , and false if the examined patient hasn't. the probability $P(s_1, s_2, \dots, s_n)$ is same as the Markov model from the last section, and the probability $P(u_1, u_2, \dots, u_n)$ is the prior probability of seeing a particular sequence of pale face events e.g. {true, false, true}. The probability $P(u_1, u_2, \dots, u_n | s_1, s_2, \dots, s_n)$ can be estimated as $\prod_{i=1}^n P(u_i | s_i)$, if all i , given s_i, u_i is independent of all u_i and s_j , for all $j \neq i$.

Suppose that the patient severity stage was in Normal stage; the examined patient had a pale face on day 2, but not on day 3. Assuming that the prior probability that the

examined patient has pale face is 0.5 then the probability that the disease severity stage is critical in day 3 can be computed as follow: -

$$\begin{aligned}
 P(s_3 = C \setminus s_1 = N, u_2 = T, u_3 = F) &= P(s_2 = \text{critical}, s_3 = \text{critical} \setminus s_1 = \\
 &\text{normal}, u_2 = \text{True}, u_3 = \text{False}) \\
 &\quad + P(s_2 = \text{medium}, s_3 = \text{critical} \setminus s_1 = \text{normal}, u_2 = \\
 &\quad \text{True}, u_3 = \text{False}) \\
 &\quad + P(s_2 = \text{normal}, s_3 = \text{critical} \setminus s_1 = \text{normal}, u_2 = \\
 &\quad \text{True}, u_3 = \text{False}) \\
 &= \frac{P(u_3=F \setminus s_3=C)P(u_2=T \setminus s_2=C)P(s_3=C \setminus s_2=C)P(s_2=C \setminus s_1=N)P(s_1=N)}{P(u_3=F)P(u_2=T)P(s_1=N)} + \\
 &\quad \frac{P(u_3=F \setminus s_3=C)P(u_2=T \setminus s_2=M)P(s_3=C \setminus s_2=M)P(s_2=C \setminus s_1=N)P(s_1=N)}{P(u_3=F)P(u_2=T)P(s_1=N)} + \\
 &\quad \frac{P(u_3=F \setminus s_3=C)P(u_2=T \setminus s_2=N)P(s_3=C \setminus s_2=N)P(s_2=N \setminus s_1=N)P(s_1=N)}{P(u_3=F)P(u_2=T)P(s_1=N)} + \\
 &= \frac{(0.7)(0.3)(0.55)(0.2)}{(0.5)(0.5)} + \frac{(0.7)(0.8)(0.2)(0.05)}{(0.5)(0.5)} + \frac{(0.7)(0.1)(0.2)(0.75)}{(0.5)(0.5)} \\
 &= 0.1568
 \end{aligned}$$

3. Chronic Kidney Disease (CKD)

Chronic kidney disease (CKD) is one of a growing health problem worldwide. It is a progressive disease associated with a high risk of cardiovascular diseases, mortality and morbidity rates, and high health care costs. Therefore, early detection of the disease to control and manage the consequences is highly significance., Because of the dynamic nature of CKD, its covert nature in the early stages, and heterogeneity of patients, predicting the CKD progression with reasonable accuracy is necessary.

CKD progression can be considered as a function of various parameters including underlying renal diseases, blood pressure, hypertension, proteinuria, and age., Early diagnosis of the CKD are takes great attention among physicians, especially on

determining the appropriate time to apply medical treatments and to control identified risk factors that reflect on the disease progression to End Stage Renal Disease (ESRD) like hypertension, proteinuria, and hyperphosphatemia.

Stages of Chronic Kidney Disease

The stages of CKD (Chronic Kidney Disease) are mainly based on measured or estimated GFR (Glomerular Filtration Rate), there are five stages, but kidney function is normal in Stage 1, and minimally reduced in Stage 2.

The KDOQI (Kidney Disease Outcomes Quality Initiative) stages of kidney disease are:

Table 4. CKD Stages according to GFR measurement value

Stage	GFR	Description	Treatment stage
1	90+	Normal kidney function but urine findings or structural abnormalities or genetic trait point to kidney disease	Observation, control of blood pressure.
2	60-89	Mildly reduced kidney function, and other findings (as for stage 1) point to kidney disease	Observation, control of blood pressure and risk factors.
3A 3B	45-59 30-44	Moderately reduced kidney function	Observation, control of blood pressure and risk factors.
4	15-29	Severely reduced kidney function	Planning for end stage renal failure.
5	<15 or on dialysis	Very severe, or end stage kidney failure (sometimes call established renal failure)	Treatment choices.

Definition of chronic: Labelling someone as having CKD requires two samples at least 90 days apart. Historical values can be used.

eGFR is estimated Glomerular Filtration Rate, usually based on serum Creatinine level, age, sex, and race. The most widely used method for this is the abbreviated MDRD (Modification of Diet in Renal Disease) equation, as it has proved the most robust and accurate. Normal GFR is approximately 100mls/min/1.73m².

$$\text{eGFR} = 186 \times (\text{Creatinine} / 88.4)^{-1.154} \times (\text{Age})^{-0.203} \times (0.742 \text{ if female}) \times (1.210 \text{ if black})$$

4. Support Vector Machine (SVM)

Data mining is the process of extracting hidden information from the large dataset. Data mining techniques are used in various applications like fault diagnosis, anomaly detection, medical diagnosis, e-mail filtering, face recognition and oil spill detection. Data mining techniques such as classification, clustering and association rule etc. plays a great role in extracting unknown knowledge from the databases. Classification is a data mining technique used to predict group membership for data instances., the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute., one of data mining Techniques is Support Vector Machine.,

SVM is a method for the classification of both linear and nonlinear data., SVM is an algorithm that works as follows. It uses a nonlinear mapping to renovate the unique training data into a higher dimension. Surrounded by this new dimension, it examines for the linear optimal separating hyperplane i.e., a “decision boundary” sorting out the tuples of one class from another. With a suitable nonlinear mapping to a necessarily high dimension, data from two classes can always be separated by a hyperplane. The SVM finds the hyperplane using support vectors and margins. Although the training time of even the fastest SVMs can be exceedingly slow, they are extremely accurate, outstanding

to their ability to model complex nonlinear decision boundaries. They are much less prone to over fitting than other methods. The Support vectors initiate also provide a compact description of the learned model. SVMs can be used for prediction along with classification. They have been applied to a several areas, including handwritten digit recognition, object recognition, and speaker identification and benchmark time-series prediction tests.

5. Conclusion

More work and research still in progress to formalize a new approach to estimate and predict chronic kidney disease stage transition rates using the Hidden Markov Model and Support Vector Machine Data Mining technique which will support physicians & policymaker in their medical & economical decisions

References

1. MacDonald, Iain L. and Zucchini, Walter (1997). Hidden Markov and Other Models for Discrete-valued Time Series. Chapman Hall.
2. MacDonald, Iain L. and Zucchini, Walter (2009). Hidden Markov Models for Time Series: An Introduction Using R. Chapman Hall.
3. Kay, Richard (1986). A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. Biometrics, Vol.42, No.4, 855-865.
4. Satten, Glen A., Longini Ira M., Jr. (1996). Markov Chains with Measurement Error: Estimating the “True” Course of a Marker of the Progression of Human Immunodeficiency Virus Disease. Applied Statistics, Vol. 45, No.3, 275-309.

5. Jackson, C., Sharples, L., Thompson, S., Duffy, S., and Couto, E. (2003). Multistate Markov Models for Disease Progression with Classification Error. *Journal of the Royal Statistical Society. Series D*, Vol. 52, No.2,193-209.
6. Shirley, K., Small, D., Lynch, K., Maisto, S., and Oslin, D. (2010). Hidden Markov Models for Alcoholism Treatment Trial Data. *Annals of Applied Statistics*, Vol. 4, No. 1, 366-395.
7. Rabiner, L. R. (1986). An Introduction to Hidden Markov Models. *IEEE Acoustics, Speech, and Signal Processing* January, 4-16.
8. Jackson, C. and Sharples, L. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant. *Statist. Med.*, 21,113-128.
9. Scott, Steven L. (1999). Bayesian Analysis of a Two-State Markov Modulated Poisson Process. *Journal of Computational and Graphical Statistics*, Vol.8, No. 3, 662-670.
10. Scott, Steven L. (2002). Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century. *Journal of the American Statistical Association*, Vol. 97, No. 457, 337-351.
11. Gentleman R.C, Lawless, J.F., Lindsey, J.C., and Yan, P. (1994). Multi-State Markov Models for Analysing Incomplete Disease History Data with Illustrations for HIV Disease. *Statistics in Medicine*, Vol. 13, 805-821.
12. Bureau, A., Hughes, J., Shiboski, S. (2000). An S-Plus Implementation of Hidden Markov Models in Continuous Time. *Journal of Computational and Graphical Statistics*, Vol. 9, No.4, 621-632.

13. Levey, AS., Bosch, JP., Lewis, JB., Greene, T., Rogers, N., Roth, D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Annals of Internal Medicine*,130, 461-470
14. Scott, S., James, G. and Sugar, C. (2005). Hidden Markov Models for Longitudinal Comparisons. *Journal of the American Statistical Association*, Vol 100, Issue 470.
15. Jackson, C. (2007). Multi-state modelling with R: the msm package. Medical Research Council Biostatistics Unit: Cambridge, UK.
16. Altman, R.M. (2007), Mixed Hidden Markov Models: An Extension of the Hidden Markov Model to the Longitudinal Data Setting, *Journal of the American Statistical Association*, 102, 201-210.
17. S. Vijayarani, S.Dhayanand ,2015 "KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS" ,*International Journal of Computing and Business Research (IJCBR)*, Vol 6, Issue 2
18. Ishani Mandli, Mahesh Panchal, (2014) "Selection of Most Relevant Features from High Dimensional Data using IG-GA Hybrid Approach", *International Journal of Computer Science and Mobile Computing*, Vol.3 Issue 2, 827-830
19. Sunitha Devi1. P, Sowjanya. CH, Sunitha. K.V.N, (2014) "A Review of Supervised Learning Based Classification for Text to Speech System", *International Journal of Application or Innovation in Engineering & Management*, Vol 3, Issue 6, 79-86
20. Sunitha Devi1. P, Sowjanya. CH, Sunitha. K.V.N, (2014) "A Review of Supervised Learning Based Classification for Text to Speech System", *International Journal of Application or Innovation in Engineering & Management*, Vol 3, Issue 6, 79-86
21. Sonali Sonavane, Alka Khade, Gaikwad. V.B, (2013) "Novel Approach for Localization of Indian Car Number Plate Recognition System using Support Vector

Machine”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 8, 179-183

22. Shaohua Teng, Hongle Du, Naiqi Wu, Wei Zhang, Jiangyi Su, (2010) A Cooperative Network Intrusion Detection Based on Fuzzy SVMs, JOURNAL OF NETWORKS, VOL. 5, NO. 4, APRIL, 475-483

23. Chronic Kidney Disease, Clinical Practice Recommendations for Primary Care Physicians and Healthcare Providers 2011, HENRY FORD HEALTH SYSTEM.

Subset Selection in Multiple Linear Regression (Review)

Ahmed Amin El-Sheikh*, Asmaa Mohamad Wahba**

Abstract:

This paper provides a brief review of subset selection criteria, in particular stepwise criterion, expected mean squared prediction error criterion, distance between predicted values based on the full model and predicted values based on the sub model criterion, Kullback's discrepancy criterion, The Loss Rank Principle (LoRP) criterion, and other Criteria.

Keywords: EMSE, Multiple Linear Regression, Stepwise, Subset Selection.

1. Introduction:

Subset selection is one of the most fundamental problems in statistical applications, it appears with the need to model the relationship between variable of interest and a set of explanatory variables, and then there is an urgent need to answer the following question; which variables will be selected to enter the model among competing explanatory variables.

Suppose Y is a variable of interest, and $X_1, \dots, X_k$ is a set of prospective explanatory variables, the problem of variable selection, or subset selection as it is often called, arises when one wants to model the relationship between Y and a subset of $X_1, \dots, X_k$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when k is large and $X_1, \dots, X_k$ is thought to contain many redundant or irrelevant variables.

The problem of subset selection lies in how to choose the variables which construct the "best" model among the alternative models. Over the decades, scientists have investigated an assortment of statistical approaches to select among alternative models (the number of possible subsets of one or more variables out of k is (2^k-1)). **Miller (1990)**

*Professor of Statistic, Department of Applied Statistic and Econometrics, Institute of Statistical Studies and Research, Cairo University, Egypt.

E-Mail: Aham103@yahoo.com

** Assistant Lecturer, Department of Mathematics, Insurance, and Statistic, Faculty of Commerce, University of Sadat city, Egypt.

E-Mail: asmaa_wahba@com.usc.edu.eg

2. Stepwise criterion:

Some of the most popular methods are forward selection, stepwise regression and backward elimination. Forward selection method starts with a model that has just a constant in it. Next, choose the variable that satisfies the criteria (e.g. R^2 , p value, actual F change, etc) that has been set for entry. The following step are to choose from remaining variable the variable that meets the entry criteria, keeping doing this until there are no variables that meet the entry criteria. On the contrary, backward elimination starts with all variables, including a constant if there is one. The variable which meets the elimination criteria is chosen for deletion. In each step, all the variable in the models have been evaluated, then remove the one that meet the elimination criteria. This process will be stopped when there are no more variables that meet the criteria for removal. **Norusis (2006)**

Stepwise regression is used to mean Efroymson's algorithm which proposed by Efroymson in 1960. This method differs from forward selection in that after each added variable to the set of selected variables, a test is made to see if any of the previously selected variables can be deleted. **Miller (1990).**

In spite of the fact that the forward selection, stepwise regression and backward elimination methods are remaining popular in many computer software packages, there are many critiques to these methods. Of which, there is no guarantee that the final model is the optimal one. In addition, the procedure reaches to a single final model, although in some cases there are several good models.

3. Expected mean squared prediction error criterion:

Some popular subset selection methods are based on estimating the predictive ability of each model. The predictive ability of the model could be measured by the expected mean squared prediction error i.e.,

$$EMSE(M_j) = E(y_{new} - \hat{y}_{new})^2$$

Where y_{new} is randomly selected new data from population and $\hat{y}_{new}$ is the prediction for y_{new} generated by model M_j .

By Estimating $EMSE$ for all possible models and choosing the model that has the smallest $EMSE$. Although it is difficult to Estimate $EMSE$ because the data about the all possible samples is not available, many studies tried to present an unbiased estimate of $EMSE$. **John, et al. (2005)**

Mallows (1973) investigated an unbiased estimate of $EMSE$ and recommended the use of C_p plots to help gauge subset selection for p explanatory variables, this statistic is defined as:

$$C_p = \left(\frac{RSS_p}{\hat{\sigma}^2} \right) - n + 2p$$

Where RSS_p is the residual sum of squares for the subset model containing p explanatory variables including the intercept (i.e., the number of parameters in the subset model), $\hat{\sigma}^2$ is an estimate of the error variance σ^2 . The mean-square error (MSE) from the full model is often used as the estimate of σ^2 . So $\hat{\sigma}^2 = MSE$ for the full model (full model means the model containing all explanatory variables of interest counting the intercept), n is the number of observations. Usually c_p is plotted against p for the collection of subset models. Acceptable models in the sense of minimizing the total bias of the predicted values are those models for which C_p approaches the value p . (i.e., those subset models that fall near the line $C_p = p$ in the plot).

Kobayashi and Sakata (1990) illustrated that subset selection using Mallows C_p is biased. This is because in some cases the probability of selecting one from two linear models is not equal. As Mallows C_p criterion gives greater selecting probability to the simpler model from two nested models equal in mean square predicted error. Kobayashi and Sakata presented a bias correction of the Mallows C_p by suggesting a critical value for statistical test. This test is

requested only in the case if the two models equals in Mallows C_p or the difference of Mallows C_p statistics is less than one in absolute value.

Ronchetti and Staudte (1994) modified a robust version of Mallows's C_p for regression models. It is defined by

$$RC_p = \frac{W_p}{\hat{\sigma}^2} - (U_p - V_p)$$

Where W_p is a weighted residual sum of squares computed from a robust fit of model with p independents, $\hat{\sigma}^2$ is a robust and consistent estimator of σ^2 in the full model, and U_p and V_p are constants depending on the weight function and the number of parameters in model p . The robust model selection procedure based on RC_p allows choosing the models that fit the majority of the data by taking into account the presence of outliers and possible departures from the normality assumption on the error distribution.

Gilmour (1996) indicated that if the MSE from the full model is used to estimate σ^2 , the distribution of $\left(\frac{RSS_p}{\hat{\sigma}^2}\right)$ can be obtained and gives an expected value of C_p which is not p . Moreover, Gilmour (1996) cited the words of Myers (1992) who said “Since the residual mean square for the complete model need not be the smallest estimate of σ^2 among those for the candidate models, it is quite possible that the equation will yield a $C_p < p$ for a few of the candidate models”. So Gilmour suggested an adjustment to C_p . The motivated criterion is called $\bar{C}_p$, an adjusted C_p has the property that if the estimate of σ^2 comes from the full mode $E(\bar{C}_p) = p$.

$$\bar{C}_p = C_p - \frac{2(k - p + 1)}{n - k - 3}$$

Where k is a total number of candidate regressors.

Sommer and Huggins (1996) suggested a criterion based on the Wald test statistic and is defined by T_p . The selected submodels is the model that have T_p -values close to or smaller than p , and, as with Mallows's C_p , they will be selected by graphical rather than stepwise methods.

$$T_p = W_p - k + 2p$$

where W_p is the Wald statistic for testing whether the coefficients of the variables not in the submodel are = 0, k and p are the numbers of parameters in the full and submodel respectively.

As T_p is equivalent to C_p in least squares linear regression, it may not be reliable in the presence of outliers. Sommer and Huggins suggested a robust version of T_p (RT_p), based on generalized M-estimator. RT_p need less computational time than RC_p because the weights are determined from the full model only. Furthermore, RT_p and T_p do not need to define residuals like C_p .

4. Distance between predicted values based on the full model $\hat{Y}_k$ and predicted values based on the sub model $\hat{Y}_p$ criterion:

Kashid and Kulkarni (2002) investigated a more general criterion called S_p criterion, for subset selection in the multiple linear regression models. S_p -criterion is unlike the subset selection methods which are based on the Least Squares (LS) Estimator. If the data contain an influential (outlier) observation or the distribution of the error variable is not normality, the LS estimator performs 'poorly' and hence a method based on this estimator (for example, Mallows' C_p -criterion) tends to select a 'wrong' subset. S_p -criterion overcomes this problem and it can be used with any type of estimator (either the LS estimator or any robust estimator).

$$S_p = \frac{\sum_{i=1}^n (\hat{Y}_{ik} - \hat{Y}_{ip})^2}{\sigma^2} - (k - 2p),$$

$\hat{Y}_{ik}$ the vector of predicted values based on the full model and $\hat{Y}_{ip}$ the vector of predicted values based on submodel, σ^2 the variance of the error.

Sp statistic is calculated for every model (for all possible values of p) and select the subset for which the corresponding value of Sp is approximately equal to p. the graph of the values of Sp against p will help in selecting a better subset.

Kashid and Kulkarni (2003) suggested a subset selection criterion in a multiple linear regression model in case if the distribution of error variable is heavy tailed. The motivated criterion called Np-criterion, and based on the nonparametric estimator of regression coefficients. It is explained that the performance of Np-criterion is better than RCp and Cp irrespective of the type of underlying distribution. Moreover Np is robust to outliers.

$$N_p = \frac{\sum_{i=1}^n (\hat{Y}_{ik} - \hat{Y}_{ip})^2}{\tau^2} - (k - 2p),$$

$\hat{Y}_{ik}$ the vector of predicted values based on the full model and $\hat{Y}_{ip}$ the vector of predicted values based on sub model, The scale parameter τ is similar to σ in the least square regression. τ measures the variability in the distribution of errors in non-parametric regression.

Dorugade and Kashid (2010 a) motivated a method called Rp-criterion. The authors used the method of ridge regression to estimate the model parameter. The study assumed that there is a problem of multicollinearity in the data. Rp- criterion is defined as:

$$R_p = \frac{\sum_{i=1}^n (\hat{Y}_{ik} - \hat{Y}_{ip})^2}{\sigma^2} - tr(\hat{H}_R \hat{H}_R) + tr(\hat{H}_{Rp} \hat{H}_{Rp}) + P$$

Where H_R and H_{RA} are the hat matrices for the the full model and the sub model respectively, and defined as follows: $H_R = X(\hat{X}X + rI)^{-1} \hat{X}$, $H_{Rp} = X_p(\hat{X}_p X_p + rI)^{-1} \hat{X}_p$, and X_p is an $n \times p$ matrix of the observations on p-1 predictors.

The procedure of R_p - criterion is beginning with computing the value of statistic for all possible subset models. Then select a subset of minimum size, for which the value of the statistic is close to 'p' or plot the values of statistic R_p vs. 'p' for all possible subset models and select the subset which is closer to the line i.e. R_p statistic = P .

Dorugade and Kashid (2010 b) proposed a new subset selection method called GR_p - criterion. They have used a various bias constant determination methods proposed by the researchers for the generalized ridge regression. GR_p - criterion is defined as:

$$GR_p = \frac{\sum_{i=1}^n (\hat{Y}_{ik} - \hat{Y}_{ip})^2}{\sigma^2} - tr(\hat{H}_{GR}H_{GR}) + tr(\hat{H}_{RA}H_{RA}) + P$$

Where, H_{GR} and H_{RA} denote the prediction matrix for the full model and subset model respectively based on the generalized ridge estimator.

One of an important advantage of the proposed method is that it can be used with least squares estimator or generalized ridge estimator of β without any modification in the proposed statistic.

Jadhav et al. (2014) developed a generalized version of S_p statistic for subset selection in the presence of outlier and multicollinearity. The performance of the proposed method is considerably better as compare to some alternative methods when the outlier observations and multicollinearity occur simultaneously in the data.

$$GS_p = \frac{\sum_{i=1}^n (\hat{Y}_{ik} - \hat{Y}_{ip})^2}{\sigma^2} - tr[(H - H_1)'(H - H_1)] + p,$$

Where, H and H_1 denote the prediction matrix for the full model and subset model respectively based on the Jackknifed Ridge M-estimator.

GS_p statistic is calculated for every model (for all possible values of p), then select the subset for which the corresponding value of GS_p is approximately equal to p .

5. Kullback's discrepancy criterion:

Some subset selection techniques depended on “Kullback's discrepancy” to measure the discrepancy between the unknown true likelihood function and a model- specific likelihood function. Kullback–Leibler (K-L) information or distance, The K-L distance has also been called the K-L discrepancy, divergence, information, and number (Kullback,1987 preferred the term discrimination information).

In terms of a probability distribution the full reality or truth model which we want to make inferences and an approximating model is noted f and g respectively. Kullback–Leibler information or distance between full reality or truth model and an approximating model is defined for continuous functions as the (usually multi-dimensional) integral.

$$I(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) dx,$$

We use the notation $f(x)$ to denote that integration is over the variable x , and use θ to represent generally a parameter or vector of parameters. Thus the notation $g(x|\theta)$ refers to the estimated models have specific parameter values from ML or LS estimation based on the given data and model. $\log$ denotes the natural logarithm. The notation $I(f, g)$ denotes the “Information lost when g is used to approximate f .” Also, $I(f, g)$ is the distance from g to f . **Burnham and Anderson (2002)**

Akaike (1974) demonstrated a procedure called Akaike information criterion AIC . AIC is an estimate of relative, expected K-L information between a fitted model and the true model. (AIC) and defined by:

$$AIC = -2 \log(L) + 2k$$

Where, L is the maximum likelihood estimates of the parameters, and k is the number of parameters within the model. AIC is calculated for a set of candidate models, and then choose the mode that has the least AIC value.

It should be noted that even if all candidate models are poor, *AIC* will select one to be the best. Thus choosing the model using *AIC* doesn't mean that the chosen model is will defined but it is the best among candidate models. Moreover *AIC* should not used to compare models from different data sets, because the inference is conditional on the given data set. *AIC* may perform poorly in small sample (when the sample size is small with respect to the number of estimated parameters). **Burnham and Anderson (2002)**

Schwarz (1978) introduced a procedure depended on Bayesian context Bayesian information criteria (BIC) (sometimes called Schwarz's information criterion (SIC)). From a Bayesian perspective, BIC is designed to find the most probable model given the data.

$$BIC = -2 \log(L) + k \log(n)$$

It is noted that BIC differs from AIC only in the second term which depends here on sample size n . Models that minimize the Bayesian Information Criteria are selected. [see Acquah (2010)]

Sugiura (1978) developed a version of *AIC* by modified the bias and obtained the criterion denoted by correct *AIC* (C- *AIC*).

$$C-AIC = (-2) \log(L) + \frac{2n(m+1)}{n-m-2}$$

Where m is the rank of the designed matrix.

Hurvich and Tsai (1989) developed a version of *AIC* that is applicable to the small sample size relation to the number of estimated parameters. They presented a small-sample bias adjustment, and introduced a modification that led to AIC_c criterion.

$$AIC_c = (-2) \log(L) + 2k \frac{n}{n-k-1}$$

Konishi and Kitagawa (1996) proposed a subset selection criteria called Generalised information criteria (GIC). GIC is derived from the information-theoretic in the context of functional statistics without the assumption that the specified family of probability distributions contains the true structure generating the data. The proposed criteria is not restricted to maximum likelihood estimation but may be applied to the evaluation of models constructed by a different of methods, unlike AIC dealing models estimated by the maximum likelihood methods.

$$GIC(X_n; \hat{G}) := (-2n) \left\{ \frac{1}{n} \sum_{i=1}^n \log f(X_i | \hat{\theta}) - \frac{1}{n} b_1(\hat{G}) \right\}$$

Where X_n is a random sample of size n from an unknown distribution $G(x)$. θ is a p -dimensional vector of unknown parameters. $b_1(\hat{G})$ is a bias estimate obtained by replacing the unknown distribution G by the empirical distribution $\hat{G}$

Bengtsson and Cavanaugh (2006) modified the Akaike information criterion and introduced AIC_i that can be used in the broad framework of the linear state-space model. The subscript i denoted to “improved”. The results indicated that AIC_i estimates is less bias than traditional AIC or corrected AIC .

$$AIC_i = -2 \log(L) + \hat{\beta}_T(k, \theta_s)$$

$\hat{\beta}_T(k, \theta_s)$ is an approximation for bias and obtained by Monte Carlo simulation using a conveniently chosen simulation parameter θ_s

Jayakumar and Sulthan (2014) explained how the multicollinearity affected estimation using least square method and how it affected the model selection decision. The authors presented multicollinearity corrected version of generalized information criteria called modified generalized information criteria ($MGIC$) to select a best model among the various candidate models. The correction was by exchange the inflated error variance by un-inflated error variance of a model. The generalized information criterion is given as:

$$MGIC = n \log(\hat{\sigma}_{INF(e)}^2) + f(n, k)$$

Where $\hat{\sigma}_{INF(e)}^2$ is the inflated estimate of population error variance which is inflated by the variance inflation factor (VIF). $f(n, k)$ is the penalty function.

6. The Loss Rank Principle (LoRP) criterion:

Hutter (2007) proposed a modern criterion called The Loss Rank Principle (LoRP). Supset selection according to LoRP is through applying the following rule:

Given a class of linear models the best model is the one with the smallest loss rank i.e. the smallest M^{best} .

$$M^{best} = \arg \min \left\{ \frac{n}{2} \log(y^T S_\alpha y) - \frac{1}{2} \log \det S_\alpha \right\},$$

Where symbol refers to both model and regression matrices, I is the identity matrix, and $S_\alpha = (I_n - M)^T (I_n - M) + \alpha I$, where M is a projection matrix and $\alpha > 0$ is a small number to be determined. (Hutter and Tran, 2010) discussed the optimality Properties of LoRP for Variable Selection.

Tran (2011) proposed a LR criterion, for selecting shrinkage parameters for variable selection purposes. a LR criterion is derived from the loss rank principle (LoRP), that was introduced by Hutter (2007) and Hutter & Tran (2010). The model that has the lowest LR criterion is chosen as a best model.

$$LR_\lambda = \frac{n}{2} \log \|y\|^2 - \frac{n}{2} \text{KL}\left(\left(\frac{df_\lambda}{n} \mid 1 - \rho_\lambda\right)\right)$$

Where λ is a certain shrinkage parameter, df_λ equal the number of non-zero coefficients, and KL is the Kullback–Leibler divergence between the Bernoulli distributions with parameters $p, q \in (0, 1)$. $\rho_\lambda := \|y - \hat{y}_{s_\lambda}\|^2 / \|y\|^2$ and $\hat{y}_{s_\lambda}$ The fitted vector under the sub model. When λ increases from 0 to infinity, the number of non-zero coefficients will be a non-increasing step function i.e. the covariates are in turn removed from the models.

7 Other Criteria:

Greenberg and Parks (1997) provided a subset selection method based on predictive approach. The least square method is used to estimate the linear regression coefficients.

The effect of including X on the predictive covariance matrix is measured by generalized variance ratio (GVR). GVR is defined as the ratio of determinants of covariance matrices when a subset of explanatory variables is included or excluded from a regression.

$$GVR = \frac{|cov(y_0|X_{01})|}{|cov(y_0|X_0)|}$$

Where $cov(y_0|X_{01})$ is the covariance matrix from the partial model. While, $cov(y_0|X_0)$ is the covariance matrix from the full model.

Andre et al. (2000) modified the robust coefficient of determination for minimum sum of absolute errors MSAE regression.

$$R = RSR / (RSR + (n - p)\hat{\sigma}/2),$$

where RSR is the reduction in the sum of absolute residuals when P-1 independent variables included in the model i.e

$$RSR = \sum_{i=1}^n |y_i - \hat{\beta}_0| - \sum_{i=1}^n |y_i - \hat{y}_i|$$

and $\hat{\sigma} = SAR/n$, where SAR is the sum of absolute residuals.

Muller and Welsh (2005) investigated a method for subset selection based on combining a robust penalized criterion and a robust conditional expected prediction loss function that is estimated using a stratified bootstrap. The model selection method is consistent under some conditions.

Sakate and Kashid (2014) proposed a model selection criterion based on deviance called the deviance-based criterion (DBC). DBC is gotten by penalizing the difference between the fitted model and the full model. Under some conditions, DBC is appeared to be a consistent model selection criterion. The selected model asymptotically equals the optimal model relating dependant and independent variables.

8. References:

1. Acqua H. (2010), "*Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship*", Journal of Development and Agricultural Economics, Vol. 2, No.1, pp.1-6.
2. Akaike H. (1974), "*A new look at the statistical model identification*", IEEE Transactions on Automatic Control, Vol. AC-19, No. 6, pp. 716-723.
3. Andre C., Elian S., Narula S. and Tavares R. (2000), "*Coefficients of determinations for variable selection in the msae regression*", Communications in Statistics - Theory and Methods, Vol.29, No.3, pp. 623-642.
4. Bengtsson T. and Cavanaugh J. (2006), "*An Improved Akaike information criterion for state - space model selection*", Computational Statistics & Data Analysis, Vol. 50, Issue 10, pp. 2635–2654.
5. Burnham K. And Anderson D. (2002), "*Model selection and multimodel inference: A practical information-theoretic approach*", 2nd edition, Springer-Verlag, New York.
6. Dorugade A. And Kashid D. (2010 a), "*Variable selection in linear regression based on ridge estimator*", Journal of Statistical Computation and Simulation, Vol. 80, No. 11, pp. 1211-1224.
7. Dorugade A. And Kashid D. (2010 b), "*Subset selection in linear regression using generalized ridge estimator*", Journal of Statistical Theory and Practice, Volume 4, No. 3, pp. 357-389.
8. Gilmour S. (1996), "*The interpretation of Mallows's Cp-statistic*", Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 45, No. 1, pp. 49-56.
9. Greenberg E. and Parks R. (1997), "*A predictive approach to model selection and multicollinearity*", Journal of Applied Econometrics, Vol. 12, No. , pp. 67-75.

10. Hurvich C. and Tsai C. (1989), "*Regression and time series model selection in small samples*", Biometrika, Vol. 76, No. 2, pp. 297-307.
11. Hutter M. (2007), "*The loss rank principle for model selection*", In Proceedings of the 20th Annual Conf. on Learning Theory (COLT'07), Vol. 4539 of LNAI (eds N. Bshouty & C. Gentile), pp.589-603, Springer, Berlin and San Diego.
12. Hutter M. and Tran M. (2010), "*Model selection with the loss rank principle*". Comput. Statist. Data Anal, Vol. 54, 1288–1306.
13. Jadhav N., Kashid D., and Kulkarni Subhash R. (2014), "*Subset selection in multiple linear regression in the presence of outlier and multicollinearity*", Statistical Methodology, Vol. 19, pp. 44-59.
14. Jayakumar D. and Sulthan A. (2014), "*Identification of multicollinearity and its effect in model selection*" Electronic Journal of Applied Statistical Analysis, Vol. 7, Issue 1, pp. 179-153.
15. John Dziak, Runze Li and Linda Collins (2005), "*Critical review and comparison of variable selection procedures for linear regression : Technical report*", Methodology Center, Penn State University.
16. Kashid D. and Kulkarni S. (2002), "*A more general criterion for subset selection in multiple linear regression*", Commun. Statist.-Theory Meth., Vol. 31, No. 5, pp. 795–811.
17. Kashid D. and Kulkarni S. (2003), "*Subset selection in multiple linear regression with heavy tailed error distribution*", Journal of Statistical Computation and Simulation, Vol. 73, No. 11, pp. 791-805.
18. Kobayashi M. and Sakata S. (1990), "*Mallows C_p criterion and unbiasedness of model selection*", Journal of Econometrics, Volume 45, Issue 3, Pages 385-395.
19. Konishi S. and Kitagawa G. (1996), "*Generalised information criteria in model selection*", Biometrika, Vol. 83, No. 4, pp. 875-890.
20. Mallows C. (1973), "*Some comments on C_p* ", Technometrics, Vol. 15 , Iss. 4, pp. 661-675.

21. Miller A. (1990), "*Subset selection in regression*", Chapman & Hall/CRC Monographs on Statistics & Applied Probability, United States of America.
22. Muller S. and Welsh A. (2005), "*Outlier robust model selection in linear regression*", Journal of the American Statistical Association, Vol. 100, No.472, pp. 1297-1310.
23. Norusis M. (2006), "*SPSS 15.0 statistical procedures companion*", Prentice Hall.
24. Ronchetti E. and Staudte R. (1994), " *A robust version of Mallows's C_p* ", Journal of the American Statistical Association, Vol. 89, No. 426, pp.550-559.
25. Sakate D. and Kashid D. (2014), "*A deviance-based criterion for model selection in GLM*", A Journal of Theoretical and Applied Statistics, Vol. 48, No. 1, pp. 34-48.
26. Sommer S. and Huggins R. (1996), "*Variables selection using the wald test and a robust C_p* " ,Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 45, No. 1, pp. 15-29.
27. Sugiura N. (1978), "*Further analysts of the data by Akaike' s information criterion and the finite corrections*", Communications in Statistics - Theory and Methods, vol.7, No.1, pp. 13-26.
28. Tran M. (2011), "*The Loss rank criterion for variable selection in linear regression analysis*", Scandinavian Journal of Statistics, Vol. 38, pp. 466–479.

Estimation of Unbalanced Panel Data

Elhoussainy A. R.¹

Ahmed A.E.²

Aya. M. A.³

Abstract Panel data set is one of the most important kinds of data sets now days because its help the researcher to look at dynamic relationships and that would not be happened with a single cross section and time series. In this paper the unbalanced panel data approach will be presented and literature review about unbalanced Panel data models will be presented.

Key words Fixed Effect, Random Effect, and Unbalanced Panel data.

1. INTRODUCTION

Panel data includes both diminutions of time and cross section. The time dimension in panel data is represented in the observations collected over time periods on each individual. So this kind of data provid rich sources of information about the economy. Missing values in Panel data sets is very common problem facing the researchers in analyzing Panel data, this problem leads to unbalanced panels. While the nature of these unbalanced panels doesn't affect applicability of many commonly used estimators (such as Fixed Effect or Random Effect estimator), but this is not true for all estimators.

2. MODEL

$$y_{it} = x_{it}\beta + \epsilon_{it}, \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, N$$

Where $\epsilon_{it} = \mu_i + u_{it}$

- y_{it} Denotes the dependent variables
- β Denotes regression coefficient
- x_{it} Denotes the independent variables
- μ_i denotes the individual error for the cross section
- u_{it} Denotes the classical error term.

The difference between Random and Fixed Effect models is:

- $E(\mu_i x_{is}) = E(u_{it} x_{is}) = 0$ (Random Effect Model)
- $E(\mu_i x_{is}) \neq 0, E(u_{it} x_{is}) = 0$ (Fixed Effect Model)

3. Review of Unbalanced Panel Data Model

In this section literature review for unbalanced panel data will be presented

3.1 Fixed Effect Models

Alev, et al (2011) developed a semi-parametric panel model to explain the trend in UK temperatures and other weather outcomes over the last century. They used the monthly averaged maximum and minimum temperatures observed at the twenty-six Meteorological Office stations. The data is an unbalanced panel. They allowed the trend to evolve in a nonparametric way so that they obtained a fuller picture of the evolution of common temperature in the medium timescale.

Jason (2012) proposed the Chamberlain projection approach, it is a powerful tool for the analysis of linear fixed-effects models, was introduced within the context of balanced panels. He proposed the Chamberlain projection approach to unbalanced panels. The extension is especially useful for models with sequential exogeneity, where existing control-variable approaches are not applicable. A generalized method of moments (GMM) estimation framework is considered, and hypothesis tests (testing strict exogeneity, testing random effects, etc.) are discussed within the GMM context.

Jianhua, et al (2012) proposed missing data in longitudinal studies can create enormous challenges in data analysis when coupled with the positive-definiteness constraint on a covariance matrix. For complete balanced data, the Cholesky decomposition of a covariance matrix makes it possible to remove the positive-definiteness constraint and use a generalized linear model setup to jointly model the mean and covariance using covariates. **Kristiaan and Ignace (2014)** explored the effect of balancing unbalanced panel data when estimating primal productivity indices using non-parametric frontier estimators. First, they listed a series of pseudo-solutions aimed at making an unbalanced panel balanced. Then, they discussed some intermediate solutions (e.g., balancing 2-years by 2 years). Furthermore, they linked this problem with a variety of literatures on infeasibilities, statistical inference of non-parametric frontier estimators, and the index theory literature focusing on the dynamics of entry and exit in industries.

Wei, et al (2014) introduced “Planned missing designs” they became increasingly popular, but because there is no consensus on how to implement them in panel data research, they simulated panel data to distinguish between strategies of assigning items to forms and of

assigning forms to participants across measurement occasions. Using relative efficiency as the criterion, results indicate that balanced item assignment coupled with assigning different forms over time most often yields the optimal assignment method, but only if variables are reliable.

3.2 Random Effect Models

Jinhong and Xian (2008) proposed the estimation of the unbalanced panel data partially linear models with a one-way error components structure. A weighted semiparametric least squares estimator (WSLSE) is developed using polynomial spline approximation and least squares. They showed that the WSLSE is asymptotically more efficient than the corresponding unweighted estimator for both parametric and nonparametric components of the model.

Boris (2009) extended the (Least Squares Dummy Variables) LSDV bias-corrected estimator in (Bun and Carree, 2005) to unbalanced panels and discusses the analytic method of obtaining the solution. Using a Monte Carlo approach, he compared the performance of this estimator with three other available techniques for dynamic panel data models.

Jeffrey (2009) proposed some strategies for allowing unobserved heterogeneity to be correlated with observed covariates and sample selection for unbalanced panels. The methods are extensions of the Chamberlain-Mundlak approach for balanced panels. Even for nonlinear models, in many cases the estimators can be implemented using standard software. The framework suggests straightforward tests of correlation between heterogeneity and the covariates, as well as sample selection that is correlated with unobserved shocks while

allowing selection to be correlated with the observed covariates and unobserved heterogeneity.

David (2010) proposed a simple analytical approach for constructing one-sided β -content, γ confidence tolerance limits is proposed for general random effects models with normal data in both balanced and unbalanced data scenarios. The approach is based on an approximation to the noncentral t distribution and modified large sample methods for constructing confidence bounds on functions of variance components. An alternative bootstrap-adjusted limit is also proposed.

Jun, et al (2010) introduced panel data with covariate measurement error appear frequently in various studies. Due to the sampling design and/or missing data, panel data are often unbalanced in the sense that panels have different sizes. For balanced panel data (i.e., panels having the same size), there exists a generalized method of moments (GMM) approach for adjusting covariate measurement error, which does not require additional validation data. They extended the GMM approach of adjusting covariate measurement error to unbalanced panel data. Two health related longitudinal surveys are used to illustrate the implementation of the proposed method.

Erik (2013) proposed a system of regression equations for analyzing panel data with random heterogeneity in intercepts and coefficients, and unbalanced panel data is considered. A maximum likelihood (ML) procedure for joint estimation of all parameters is described. Since, its implementation for numerical computation is complicated, simplified procedures are presented. The simplifications essentially concern the estimation of the covariance matrices of the random coefficients. The application and 'anatomy' of the proposed algorithm for modified ML

estimation are illustrated by using panel data for output, inputs and costs for 111 manufacturing firms observed up to 22 years.

Wei and Lung (2013) suggested and compared different methods for estimating spatial autoregressive panel models with randomly missing data in the dependent variable. They started with a random effects model and then generalize the model by introducing the spatial Mundlak approach. A nonlinear least squares method is suggested and a generalized method of moments estimation is developed for the model. A two-stage least squares estimation with imputation is proposed as well. They analytically compared these estimation methods and find that the generalized nonlinear least squares, best generalized two-stage least squares with imputation, and best method of moments estimators have identical asymptotic variances.

Liwen and Dengkui (2016) presented parametric bootstrap (PB) approaches for hypothesis testing and interval estimation for the regression coefficients of panel data regression models with incomplete panels. Some simulation results are presented to compare the performance of the PB approaches with the approximate inferences. They showed that the PB approaches perform satisfactorily for various sample sizes and parameter configurations, and the performance of PB approaches is mostly better than the approximate methods with respect to the coverage probabilities and the type I error rates.

References

1. Alev, A., Oliver, L. and Zhijie X. (2011). A semiparametric panel model for unbalanced data with application to climate change in the United Kingdom. *Journal of Econometrics*. 164, 92 – 115.
2. Boris, L. (2009). Further results on bias in dynamic unbalanced panel data models with an application to firm R&D investment. *Applied Economics Letters*.1227-1233.
3. Erik, B. (2013). Estimating SUR system with random coefficients: the unbalanced panel data case. *Empir Econ*. 47, 451–468.
4. Jason, A. (2012). The projection approach for unbalanced panel data. *Econometrics Journal*. 16,161–178.
5. Jeffrey, M. (2009). Correlated Random Effects Models With Unbalanced Panels. Department of Economics, Michigan State University. Presented at the 15th Conference on Panel Data, Bonn, Germany, July 3-5, 2009.
6. Jianhua, Z., Min,C., Mehdi,M. and Mohsen,P. (2012). A cautionary note on generalized linear models for covariance of unbalanced longitudinal data. *Journal of Statistical Planning and Inference*.142, 734-751.
7. Jinhong, Y. and Xian, Z. (2008). Partially linear models and polynomial spline approximations for the analysis of unbalanced panel data. *Journal of Statistical Planning and Inference*139, 679-695.
8. Jun,S., Zhiguo, X. and Ruifeng ,X. (2010). Estimation with unbalanced panel data having covariate measurement error. *Journal of Statistical Planning and Inference*. 141, 800-808
9. Kristiaan, K. and Ignace, V. (2014). Comparing Malmquist and Hicks–Moorsteen productivity indices: Exploring the impact of unbalanced

- vs. balanced panel data. European Journal of Operational Research.233, 749–758.
10. Wei, W. and Lung, F. (2013). Estimation of spatial panel data models with randomly missing data in the dependent variable. Regional Science and Urban Economics. 43, 521-538.
 11. Wei, W., Terrence, D., Mijke, R., Alexander, S., Brent, M., and Todd, D. (2014). Optimal assignment methods in three-form planned missing data designs for longitudinal panel studies. International Journal of Behavioral Development. 38(5), 397–410.

Websites

12. David, H. (2010). One-Sided Tolerance Limits for Balanced and Unbalanced Random Effects Models. 303-312. Taylor and Francis Group.
(http://amstat.tandfonline.com/doi/abs/10.1198/TECH.2010.09171#.WinA_FWGPV)
13. Liwen, X. and Dengue, W. (2016). Parametric bootstrap inferences for unbalanced panel data models. Communications in Statistics - Simulation and Computation. Published online: 24 Oct 2016
(<http://www.tandfonline.com/doi/abs/10.1080/03610918.2016.1248567?journalCode=lssp20>)

Extracting the Wealth of Knowledge in Composite Index Data

Asmaa Mohamed Sayed¹

Abstract:

Composite indices data are high-dimensional data, where the number of variables sometimes exceeds 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 can not 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, Multidimensional Scaling) to develop new alternative indices for the same data.

We illustrate these methods using the Ibrahim Index of African Governance (IIAG) data for 2016, the most recent year for which the data are available at the time of writing. These data are provided by 34 well known and trusted providers. The data are collected on 54 African countries including Egypt. The data consists of 154 variables, some of which are grouped into indicators. The total number of indicators is 95. These are grouped into 4 main categories, and 14 subcategories. The main purpose is to compute an index number for each of the African countries that can be used to assess the overall performance of 54 African countries in terms of their abilities to deliver public goods and services.

Key Words:

Bootstrap; Cluster Analysis; Composite Indices; Hierarchical Clustering; Indicator Weights; Imputation; K-means Algorithm; Margin of Error; Missing Observations; Multidimensional Scaling; Principal Components Index.

¹M.Sc. Candidate, Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University.

1 Introduction

A composite index is a combination of various sources of information known as variables and indicators, measured in or of a system in order to provide a summary of the system that is itself not directly measurable. Composite indices have venerable history. Budro (2008) provides a survey of the current composite indices around the world. At that time Budro (2008) found 187 indices. Examples of composite indices are: the corruption perception index (CPI), Lambsdorff (2012) Modified corruption Index (MCPI), The Human Development Index (HDI), De Muro, Mazziotta, and Pareto (2011) proposal of modification of Human Development Index (MHDI), the Ibrahim Index of African Governance (IIAG), the Gender Inequality Index (GII), the Climate Change performance Index (CCPI), and The consumer Price Index (CPI). to mention only a few.

Most recently, the International Knowledge Index (IKI) was computed for the first time and published in November 2017 by the United Nations Development Program (UNDP). The IKI extended the Arab Knowledge Index (AKI) which was computed for the first time in 2015 by the Al Maktoum Foundation (<http://www.mbrf.ae/>) to measure knowledge in the Arab countries, the Happiness Index (among youth 18–29 years) by Abo Elazm (2017), and Child Wellbeing Index by Rabea (2017).

The construction of a composite index usually starts with defining the concept of interest that needs to be measured. For example, corruption as in the CPI, gender inequality as in the GII, governance as in the IIAG, or Knowledge as in the AKI and IKI. 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 we will use the terms variables and indicators synonymously.

The variables and/or indicators are then partitioned into groups (categories), subgroups, and sometimes subsubgroups. Although these groupings are obtained after extensive discussion among various experts in the field, experts do not always agree on these groupings. Even when they agree, the question is what weights should be given to indicators within subgroups, to subgroups within a group, and to groups

within the overall index. In some indices, the weights are given by experts or derived at by other means. In other indices the weights of indicators within a subgroup are given equal weight, the subgroups within a group are given equal weight, and finally the groups are given equal weight to obtain the final composite index.

For example, the IIAG consists of measurements on 54 African countries. For each country, 154 constituent variables, taken from 34 data providers, are then grouped into 95 indicators. These 95 indicators are divided into four overarching categories. Each of the four categories consists of subcategories. The total number of subcategories is 14. All indicators within a subcategory are given equal weights, all subcategories within a category are given equal weights, and finally all categories are given equal weights to reach the final IIAG.

There are several drawbacks of such construction of indices. The indicator weights depend on how one divides the indicators into groups and subgroups. The indicator weights will not be equal unless all groups have equal number of subgroups and all subgroups have equal number of indicators. Second, although the division of indicators into groups and subgroups are usually arrived at after careful consideration by various experts in the field, other partitioning of the indicators can be reasonably obtained, yielding possibly different indices of the same data. Third, it seems that this choice of weighting scheme is based on simplicity but not on any optimality criteria. We propose other complementary ways of constructing indices (scores) derived from composite index data that do not depend partitioning the indicators or a priori choice of the weights.

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).

The construction of composite indices are conceptually easy, but in practice several challenges are encountered. Perhaps, the most important one of which is data quality because the quality of an index cannot exceed the quality of the data that are used in its construction. Some of these challenges are discussed in the following sections.

1.1 Data Availability

It goes without saying that a good index should use good and well-known data providers. The data sources can provide raw or modified data. Variables should be selected based on their relevance, accessibility, and timeliness. Proxy measures can be used if the desired variables are not available. The accuracy of proxy measures can be investigated through sensitivity analysis and correlation. Their consistency and reliability can be investigated, for example, by the use of Cronbach's α (Cronbach, 1951). Sometimes practitioners tend to use whatever is available, but poor data quality will produce a poor index. A good composite index should allow improvement and development of its method of construction and renew its data over time.

1.2 Indicator Scales

Because the data come from different sources and in different units of measurements, normalization is required before any aggregation. Three of the common scaling methods mentioned in the OECD 2008 are:

1. Standardization: This method subtracts the mean of each indicator x from each score x_i then divides by the standard deviation of the scores of each indicator, that is,

$$z_i = \frac{x_i - \bar{x}}{s_x}, \quad i = 1, \dots, n, \quad (1)$$

where x_i is the i th value of the indicator x , $\bar{x}$ and s_x are the mean and standard deviation of the indicator x , and n is the number of observations. The resultant standardized indicator z will have a mean of 0 and standard deviation 1.

2. Min-Max Normalization: This method converts the score of an indicator X by subtracting the minimum from each score then dividing by the range of the scores, that is,

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \quad i = 1, \dots, n. \quad (2)$$

The values of the resultant normalized indicator y will have values from 0 to 1.

The main disadvantage of these methods, however, is that they are not robust, that is, they are affected by the presence of outliers. Of course, this disadvantage can be avoided by replacing the mean and standard deviation but more robust measures of location and scale.

3. Ranking: The scores in an indicator are replaced by their ranks. Ranking is perhaps the simplest method of normalization and it is not affected by outliers. This method allows performances of countries to be compared to each other or a country to be compared to itself based on its position or rank over time.

1.3 The Presence of Missing Values

Composite index data often contain missing values. The presence of missing values threatens the validity of the statistical inference. Missing values can be categorized as follows (Donders et al. (2006)):

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 depend on the values themselves. Like people of high income tends not to report their income.

There are various methods of handling missing data. The selection of a method is related to the type of the missing observations. The methods of handling missing values include:

1. Case Deletion: Exclude any incomplete observation from the analysis. In this case we end up with less information and the standard error of the index will increase in the reduced sample.
2. Single Imputation: Replace each missing value by newly imputed value. For example, a missing value can be replaced by the mean or the median, or by other methods such as regression imputation and expectation-maximization imputation. Single imputation underestimates the variance, because it partially reflects the uncertainty in the imputation.

3. Multiple Imputations: Replace each missing value by several newly imputed values. 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.

1.4 The Presence of Outliers

Outliers are a minority of observations that are inconsistent with the pattern suggested by the majority of observations in a data. Outliers usually lie an abnormal distances from the other observations in the data. There are several methods for the identification of outliers in composite index data. Outliers can be identified for each individual indicator or by considering indicators jointly as a multivariate data.

An example of the identification of outliers in an individual indicator is described as follows: The $\alpha\%$ trimmed mean and trimmed standard deviation are computed based on the central $(100 - \alpha)\%$ of the values. Commonly used choices of α are 5% and 10%. Then *all* observations more than 3 trimmed standard deviations away from the trimmed mean are then replaced by:

$$\text{Trimmed mean} + 3.1 \times \text{Trimmed standard deviation}, \quad (3)$$

if they are in the right tail, or by

$$\text{Trimmed mean} - 3.1 \times \text{Trimmed standard deviation}, \quad (4)$$

if they are in the left tail.

Because this rule is simple and it strikes a balance between efficiency and robustness, it is adopted by well-established indices, such as the Ibrahim Index of African Governance (IIAG) (see the MIF foundation Web site at: mo.ibrahim.foundation).

Identifying outliers in individual indicators may not be able to detect multivariate outliers, that is, outliers in the multivariate space. There are several outlier identification methods for multivariate data. See, for example, Rocke and Woodruff (1996), Rousseuw and Van Driessche (1999), and Billor et al. (2000).

1.5 The Choice of Indicator Weights

Weights can have a significant effect on the overall results of a composite index. A number of weighting schemes that exist in the literature include the following (see, e.g., Dobbie and Dail, 2013):

1. Equal Weights. This approach is commonly used because of its simplicity. It is based on the hypothesis that all indicators are equally important.
2. Data-Driven Weights. This approach is adaptive, that is, the weights are chosen depending on the data. Some of the methods of adaptive weights include (see, e.g., Decancq and Lugo, 2013):
 - (a) Frequency based weights. The choice of weights is based on the proportional of population that suffering deprivation in an indicator. For example the smaller proportion of individuals that suffer from a certain deprivation the higher weight. Or on the other hand giving less weight to those variables where problems exist.
 - (b) Expert Opinions. Experts in the domain field may be asked about their assessments of the relative weights for the indicators. A summary measure of their weights (e.g., the mean) can be used as the weights of indicators in the construction of the index.

1.6 The Margins of Errors

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, completeness of data sources, the use of imputation of missing values, etc. Therefore, index numbers should be accompanied by margins of errors or confidence bands that reflect their imprecision. This is important because the index values can be used to rank the countries. But is a given country's rank really different from other countries with neighboring ranks?

To answer this question, we need to compute the standard error of the index values that can be used to construct confidence intervals for the index values. From these intervals we can infer whether the rank of a country is significantly different from

the ranks of the other countries. The problem here is that the standard errors of the index values are not theoretically available. A way out of this is to use the Bootstrap method (Efron and Tibshirani, 1993) to estimate the standard errors. This is used, for example, by the Ibrahim Index of African Governance (IIAG).

The Bootstrap as used in the IIAG is described as follows:

1. Draw a large number M (e.g., 1,000) bootstrap samples from the IIAG data (95 indicators in 2016) within their respective subcategories (14 in 2016).
2. For each bootstrap sample compute the IIAG for the subcategories, the categories, and the over all index.
3. Compute the standard error, $SE(M_{ti})$, of the M scores, where M_{ti} is the IIAG score for a subcategory, a category, or the overall index.
4. Compute the $(1 - \alpha)100\%$ confidence bands for the expected value $\mathbb{E}(M_{ti})$ by

$$M_{ti} \pm z_{\alpha/2} SE(M_{ti}), \quad (5)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

The overall Ibrahim Index of African Governance (IIAG), M_{ti} , for 2016 together with the MIF bootstrap standard errors SE are shown in Table 1.

The confidence intervals in Table 1 can be used to determine if there are significance differences among country scores and/or among their ranks. For example, for the five highest ranked countries (1 = Mauritius, 2 = Botswana, 3 = Cabo Verde, 4 = Seychelles, and 5 = Namibia), one can see that Mauritius has a significantly higher score and rank than the other four countries, whereas the other four countries are not significantly different from each other.

One advantage of the bootstrap methodology for constructing the error bands is that it is nonparametric, that is, it does not make assumptions about the probability distribution of the indicators. However, this methodology has its own disadvantages.

1. It is computationally intensive because it requires a huge numbers of bootstrap samples for the resulting error bands to be reliable.
2. The resulting error bands are random, that is, it gives different error bands every time the method is applied to the same data.

Table 1: The 2016 IIAG and the Bootstrap Standard Errors and 95% Conf. Limits.

No.	ISO	Country	IIAG	Rank	SE	Low Limit	High Limit
1	DZ	Algeria	53.8	20	2.3	49.3	58.3
2	AO	Angola	39.2	45	2.5	34.4	44.0
3	BJ	Benin	57.5	16	1.9	53.8	61.2
4	BW	Botswana	73.7	2	2.2	69.4	78.0
5	BF	Burkina Faso	51.8	23	2.3	47.3	56.3
6	BI	Burundi	41.9	43	2.2	37.6	46.2
7	CV	Cabo Verde	73.0	3	2.1	68.8	77.2
8	CM	Cameroon	45.7	38	2.1	41.6	49.8
9	CF	Central African Republic	25.7	52	2.4	21.0	30.4
10	TD	Chad	34.8	48	2.4	30.2	39.4
11	KM	Comoros	50.3	26	2.6	45.1	55.5
12	CD	Congo	43.0	42	2.5	38.1	47.9
13	CI	Cote d'Ivoire	52.3	21	1.6	49.1	55.5
14	CG	Democratic Republic of Congo	35.8	46	2.5	30.8	40.8
15	DJ	Djibouti	46.5	37	2.7	41.1	51.9
16	EG	Egypt	51.0	24	2.5	46.1	55.9
17	GQ	Equatorial Guinea	35.4	47	2.9	29.7	41.1
18	ER	Eritrea	30.0	50	2.5	25.1	34.9
19	ET	Ethiopia	49.1	31	2.2	44.7	53.5
20	GA	Gabon	48.8	32	2.3	44.2	53.4
21	GM	Gambia	46.6	35	2.4	41.8	51.4
22	GH	Ghana	63.9	8	1.9	60.2	67.6
23	GN	Guinea	43.3	41	2.0	39.4	47.2
24	GW	Guinea-Bissau	41.3	44	2.6	36.2	46.4
25	KE	Kenya	58.9	12	2.0	55.0	62.8
26	LS	Lesotho	57.8	15	2.0	53.9	61.7
27	LR	Liberia	50.0	28	2.1	45.8	54.2
28	LY	Libya	29.0	51	2.9	23.3	34.7
29	MG	Madagascar	48.5	34	1.9	44.8	52.2
30	MW	Malawi	56.6	17	1.8	53.2	60.0
31	ML	Mali	50.6	25	1.8	47.0	54.2
32	MR	Mauritania	43.5	40	2.5	38.7	48.3
33	MU	Mauritius	79.9	1	1.9	76.3	83.5
34	MA	Morocco	58.3	14	2.3	53.8	62.8
35	MZ	Mozambique	52.3	22	1.8	48.9	55.7
36	NA	Namibia	69.8	5	1.8	66.2	73.4
37	NE	Niger	50.2	27	2.2	45.8	54.6
38	NG	Nigeria	46.5	36	2.1	42.3	50.7
39	RW	Rwanda	62.3	9	2.5	57.5	67.1
40	ST	Sao Tome and Principe	60.5	11	2.4	55.8	65.2
41	SN	Senegal	60.8	10	2.0	56.9	64.7
42	SC	Seychelles	72.6	4	2.2	68.3	76.9
43	SL	Sierra Leone	49.4	30	1.9	45.6	53.2
44	SO	Somalia	10.6	54	2.1	6.60	14.6
45	ZA	South Africa	69.4	6	2.1	65.4	73.4
46	SS	South Sudan	18.6	53	2.2	14.3	22.9
47	SD	Sudan	30.4	49	2.2	26.0	34.8
48	SZ	Swaziland	49.7	29	2.0	45.9	53.5
49	TZ	Tanzania	56.5	18	2.0	52.6	60.4
50	TG	Togo	48.5	33	2.1	44.4	52.6
51	TN	Tunisia	65.4	7	2.3	60.9	69.9
52	UG	Uganda	56.2	19	1.9	52.5	59.9
53	ZM	Zambia	58.8	13	1.7	55.5	62.1
54	ZW	Zimbabwe	44.3	39	2.1	40.2	48.4

3. It is done on the subcategory level, where the number of values to bootstrap is as small as 5 or even less.
4. The bootstrap assumes that the indicators are interchangeable within their respective subcategories.

These disadvantages cast doubt as to the appropriateness and reliability of using the bootstrap for estimating the margins of errors. One objective of this paper is to propose methods that do not depend on the computation of the margins of errors (Section 3).

In this paper, we focus our attention on three challenging issues: (a) The high dimensionality of the data, (b) The choice of weights for the indicators, and (c) Measuring the margins of errors of the index.

The rest of this paper is organized as follows: Section 2 examines alternative ways of constructing composite indices that do not depend on partitioning of the indicators. In particular, Section 2.1 considers an index which gives equal weights to all indicators. Section 2.2 proposes an index based on the Principal Components. An important advantage of these indices is that they do not depend on, nor do they require any particular partitioning of indicators into groups or subgroups.

Section 3 proposes ways of clustering the multivariate composite indices data. We can avoid both the issues of grouping variables/indicators and of the need to compute the standard errors by using clustering as explained in Section 3. We illustrate these methods using the Ibrahim Index of African Governance and the HDS2014 Children Well-Being data. This Demographic and Health Survey for Egypt is used to create measure the deprivations of the Egyptian children. The data contain three groups: Children under five years old, children aged 5 to 17 years old, and all children aged 0 to 17. The first group, contains seven deprivation indicators these indicators are: Education, Shelter, Health, Water, Sanitation, Knowledge and Nutrition. for the other two groups, the Nutrition deprivation indicator is dropped from the dominance (for details see, Abd ElAziz, 2017). Finally, summary and concluding remarks are given in Section 4.

2 Alternative Ways of Constructing Composite Indices

As alternative to the composite indices which are based on partitioning the indicators into groups and subgroups, we consider the equal-weight index (Section 2.1), where all indicators are given equal weights without grouping them. We also propose another index based on the Principal Components (Section 2.2). We then compare these indices with the classical way of index construction using the African Governance data.

2.1 Equal-Weights Index (EWI)

Composite indices are constructed by first, grouping the variables/indicators into groups, subgroups, and sometimes subsubgroups. Even within a subgroup, one may find indicators each of which is a composite index of several variables or indicators. Although these groupings are obtained after extensive discussion among the experts that are involved in the construction of the index, experts do not always agree on these groupings. Even when they agree, the question is what weights should be given to indicators within subgroups and to subgroups within a group. In some indices, the weights are given by experts or derived at by other means. In other indices the weights of indicators within a subgroup are given equal weight, the subgroups within a group are given equal weight, and finally the groups are given equal weight to obtain the final composite index.

To see the impact of different grouping and weighting schemes on the computation of the index, we consider two ways of computing the IIAG. The first is the way used by the Ibrahim Foundation, where all indicators within each sub-category are given equal weights, all sub-categories within each category are given equal weights, and all categories are given equal weights to arrive at the overall IIAG. As pointed out earlier, one advantage of this choice of “equal” weighting seems to be its simplicity, although in actuality, the final indicator weights are not equal but very different. This is because the number of indicators are not equal across subcategories and the number of subcategories are not equal across categories.

The second way is simply to ignore the groupings and subgrouping of variables and indicators and treat the variable generically by giving them equal weights. Specif-

ically, let $\mathbf{X}$ the data matrix containing all indicators as columns and all observations as rows. The Equal-Weights index (EWI) assigns equal weights to all indicators (without grouping) and the index in this case will simply be the arithmetic mean of the rows of the data matrix $\mathbf{X}$. In the case where the i th observation contain missing values, they can be ignored or replaced by the average of that row (or any other desired imputed value). Note that the EWI here is different from the indices based on equal weights which are assigned after grouping the indicators into groups and subgroups. The EWI does not partition the indicators.

The EWI is simple to compute and explain. Another important advantage of EWI is that it does not depend on, nor does it require, any particular partitioning of indicators into groups or subgroups as is the case for many composite indices. One obvious drawback of the EWI is that the choice of truly equal weights seem to be selected for simplicity.

To see the impact on grouping and weights on the IIAG index, Table 2 shows the IIAG rankings of the 54 countries in 2016, the rankings based on equal weights, and the differences between the two rankings. ranking versus the IIAG ranking of the 54 countries in 2016. Figure 1 also shows the scatter plot of equal weight ranking versus the IIAG ranking of the 54 countries in 2016. One can see from Figure 1 that the two methods are strongly related but there are substantial differences among the two rankings of the countries. For example, as can be seen in Table 2, the IIAG ranking of Swaziland is 29 but the equal weights ranking is 22, a difference of 7 positions.

2.2 Index Based on the Principal Components (PCI)

One way to deal with the high dimensionality of the data is to use multivariate dimension reduction techniques. A common technique for dimension reduction is the principal components analysis (PCA). In PCA we find an $n \times p$ matrix $\mathbf{Y} = \mathbf{XV}$, where $\mathbf{V}$ is chosen such that the variables in $\mathbf{Y}$ are orthogonal to each other and have decreasing variances. The variables in $\mathbf{Y}$, which are linear functions of the variables in $\mathbf{X}$, are called the principal components (PCs) of $\mathbf{X}$.

To find the PCs, let $\mathbf{S}$ be the variance-covariance matrix of $\mathbf{X}$ and $\lambda_1 \geq \dots \geq \lambda_p$ be the ordered eigenvalues of $\mathbf{S}$ and $\mathbf{V} = (V_1, \dots, V_p)$ be the matrix containing the corresponding eigenvectors. It can be shown that the PCs of $\mathbf{X}$ are given by

Table 2: The IIAG and Equal Weight Rankings of the 54 Countries in 2016.

No.	ISO	Country	IIAG	EWI	Difference
1	DZ	Algeria	20	20	0
2	AO	Angola	45	44	1
3	BJ	Benin	16	16	0
4	BW	Botswana	2	2	0
5	BF	Burkina Faso	23	24	1
6	BI	Burundi	43	43	0
7	CV	Cabo Verde	3	3	0
8	CM	Cameroon	38	37	1
9	CF	Central African Republic	52	52	0
10	TD	Chad	48	48	0
11	KM	Comoros	26	29	3
12	CD	Congo	42	41	1
13	CI	Côte d'Ivoire	21	25	4
14	CG	Democratic Republic of Congo	46	47	−1
15	DJ	Djibouti	37	35	2
16	EG	Egypt	24	21	3
17	GQ	Equatorial Guinea	47	46	1
18	ER	Eritrea	50	49	1
19	ET	Ethiopia	31	26	5
20	GA	Gabon	32	30	2
21	GM	Gambia	35	34	1
22	GH	Ghana	8	9	1
23	GN	Guinea	41	42	1
24	GW	Guinea-Bissau	44	45	1
25	KE	Kenya	12	13	1
26	LS	Lesotho	15	15	0
27	LR	Liberia	28	31	3
28	LY	Libya	51	51	0
29	MG	Madagascar	34	36	2
30	MW	Malawi	17	19	2
31	ML	Mali	25	27	2
32	MR	Mauritania	40	40	0
33	MU	Mauritius	1	1	0
34	MA	Morocco	14	10	4
35	MZ	Mozambique	22	23	1
36	NA	Namibia	5	6	1
37	NE	Niger	27	28	1
38	NG	Nigeria	36	38	2
39	RW	Rwanda	9	7	2
40	ST	Sao Tome and Principe	11	12	1
41	SN	Senegal	10	11	1
42	SC	Seychelles	4	4	0
43	SL	Sierra Leone	30	33	3
44	SO	Somalia	54	54	0
45	ZA	South Africa	6	5	1
46	SS	South Sudan	53	53	0
47	SD	Sudan	49	50	1
48	SZ	Swaziland	29	22	7
49	TZ	Tanzania	18	17	1
50	TG	Togo	33	32	1
51	TN	Tunisia	7	8	1
52	UG	Uganda	19	18	1
53	ZM	Zambia	13	14	1
54	ZW	Zimbabwe	39	39	0

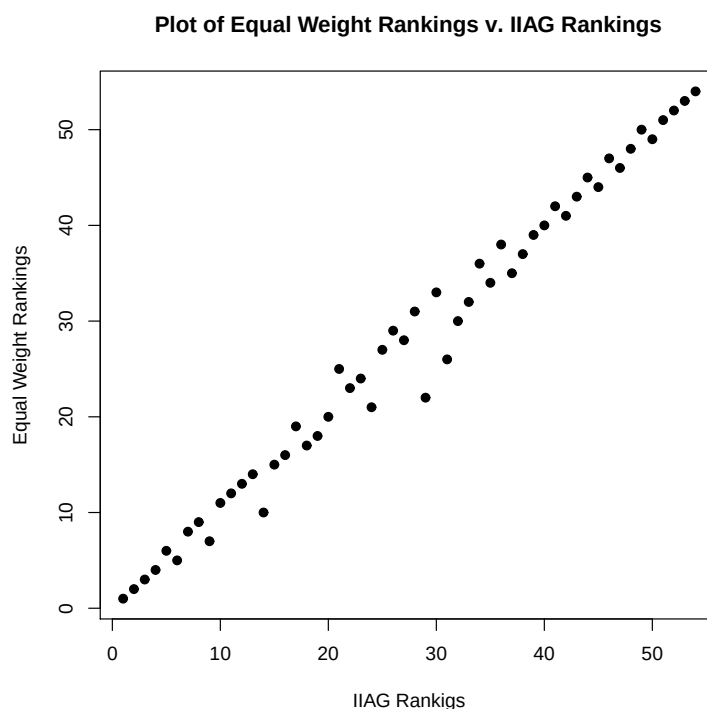


Figure 1: Plot of Equal Weight Rankings v. IIAG Rankings.

$\mathbf{Y} = \mathbf{XV}$. Furthermore, the variance of the j -th column of $\mathbf{Y}$ is the j -th eigenvalue $\lambda_j, j = 1, \dots, p$. If $p > n$, as is the case here, it can also be shown that the last $p - n = 95 - 54 = 41$ eigenvalues are zeros, which means that each of the last 41 PCs have zero variance. Accordingly, the total variance of the variables in $\mathbf{X}$ is equal to the total variance of the first 54 variables in $\mathbf{Y}$. Thus we can use the first 54 variables in $\mathbf{Y}$ instead of $\mathbf{X}$ without loss in the variance. Further reduction is possible if the variables in $\mathbf{X}$ are highly correlated but with some loss of the variance. For example if we decide to keep only $k \ll 54$ PCs, then the total variance of the first k variables in $\mathbf{Y}$ accounts for

$$\frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^{54} \lambda_j} \times 100\% \quad (6)$$

of the total variance in $\mathbf{X}$. For example for the 2016 data, the first $k = 2$ components account for 44% of the total variance, whereas the first $k = 5$ components account for 66%, the first $k = 20$ components account for 98%, and the first $k = 24$ components account for 100% (this is because $\lambda_{25} = 0$).

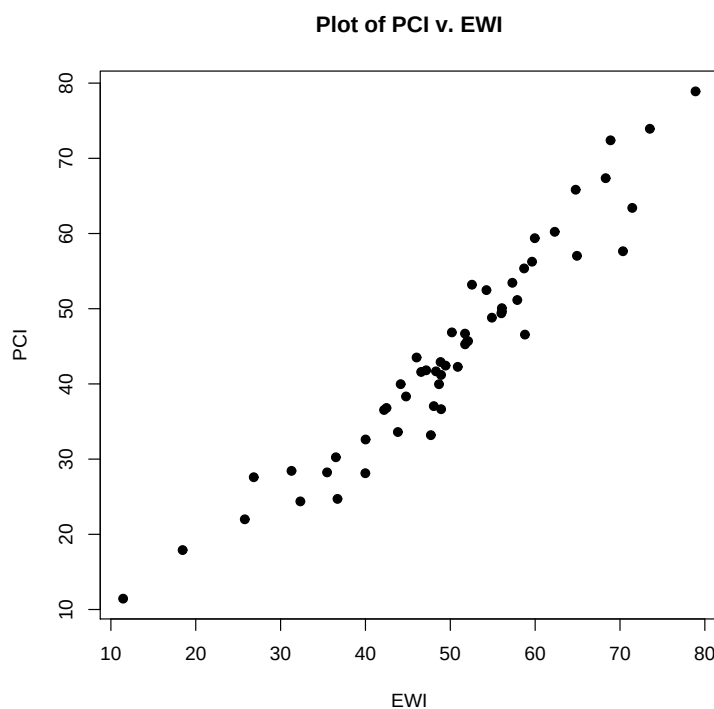


Figure 2: Scatter Plot of CPI v. EWI of the 2016 African Governance Data.

The steps for computing an index based on the Principal Components (PCI) are:

1. Compute the ordered eigenvalues $(\lambda_1, \dots, \lambda_p)$ and corresponding eigenvectors $\mathbf{V} = (V_1, \dots, V_p)$ of the matrix $\mathbf{X}^T \mathbf{X}$.
2. Set $\boldsymbol{\alpha} = \mathbf{V}_1$ and compute $\mathbf{U} = \mathbf{X}\boldsymbol{\alpha}$. If the correlation between $\mathbf{U}$ and EWI is negative, then replace $\mathbf{U}$ by $-\mathbf{U}$.
3. The PCI scores can then be transformed to an index like measure by

$$\text{CPI}_i = \min(\text{EWI}) + \frac{U_i - \min(\mathbf{U})}{\text{Range}(\mathbf{U})} \times \text{Range}(\text{EWI}), \quad (7)$$

so that CPI will have the same range as the EWI.

4. Rank countries according to CPI.

Figure 2 shows the strong linear relationship between PCI and EWI.

3 Clustering Multivariate Composite Indices Data

Instead of computing a score and a rank of each country, we can group countries into K groups such that the countries in one group are similar to each other but different from the countries in other groups. Clustering methods can be used to achieve this objective. There are numerous methods for clustering in the literature. Here we consider two of the commonly used ones in practice: the K-Means Clustering and the Hierarchical Clustering.

3.0.1 K-Means Clustering

One of the most commonly used clustering algorithm is the K-Means algorithm. It can be described briefly as follows (Johnson and Wichern, 1992):

1. Choose the number of clusters, K .
2. Generate K random points and use them as cluster centers.
3. Assign each point to the nearest cluster center.
4. Compute the new cluster centers.
5. Repeat Steps 3 and 4 until some convergence criterion is met (usually that the assignment hasn't changed).

The K -Means algorithm is popular due to its simplicity and speed (it is suitable for large data sets). But it has serious disadvantage. One needs to specify the desired number of clusters K . As a rule of thumb K is approximately $\sqrt{n/2}$. More seriously, from Step 2, the output is random, that is, it does not yield the same results with each run. Also, it tries to minimize intra-cluster variance, but does not ensure a global minimum. For these reasons we will consider here the hierarchical clustering algorithm.

3.0.2 Hierarchical Clustering

There are several algorithms that fall under the class of hierarchical clustering algorithms. One commonly used hierarchical clustering algorithms is the Agglomerative method, which is briefly summarized as follows (Rousseeuw, 1987):

1. Compute an appropriate distance (e.g., Euclidean distance) between every pair of observations.
2. Assign each of the n observations to its own cluster.
3. Compute an appropriate distance (e.g., using Ward Method or Complete Linkage distance) between every pair of clusters and merge the two closest clusters into one cluster.
4. Repeat Step 3 until all n observations belong to one cluster.
5. Draw the Dendrogram (cluster tree) and from the tree form the K clusters.

Using the Agglomerative Hierarchical Clustering method on the 2016 African Governance data without grouping, we obtain the dendrogram shown in Figure 3. From the dendrogram, one can conclude that the appropriate number of clusters (groups) is 5. The five clusters (groups) are shown within the five rectangular boxes. The countries in each of the 5 groups are given in Table 3.

Figure 4 shows the scatter plot of the Hierarchical Cluster grouping versus the IIAG rankings of the 54 countries in 2016. The plot shows some similarities but also some substantially different rankings of the countries. The top seven countries according to the IIAG are all in the Group 1 (the highest ranked group). Similarly, the four lowest ranked countries according to the IIAG rankings are all in the lowest ranked Group 5. However, Sudan (SD) with rank 49 was placed in Group 5 whereas Eritrea (ER) with rank 50 was placed in Group 4. Also, Egypt (EG) with rank 24 was placed in Group 2 surpassing many countries with better IIAG rankings. Perhaps the most surprising result is that Sao Tome and Principe (ST), which is the eleventh top ranked country by IIAG, is placed in the second lowest Group 4.

Figure 5 compares the Equal Weight rankings (see Table 2) with the Hierarchical Cluster rankings. We see here also that there are some similarities and some differences. The six top ranked countries are placed in the highest Hierarchical Cluster Group 1. Similarly, the five lowest ranked countries are placed in the lowest Hierarchical Cluster Group 5. But there are many differences. For example, Tunisia (TN), which is ranked 8th, is placed in Group 1 ahead of Rwanda (RW), which is ranked 7th.

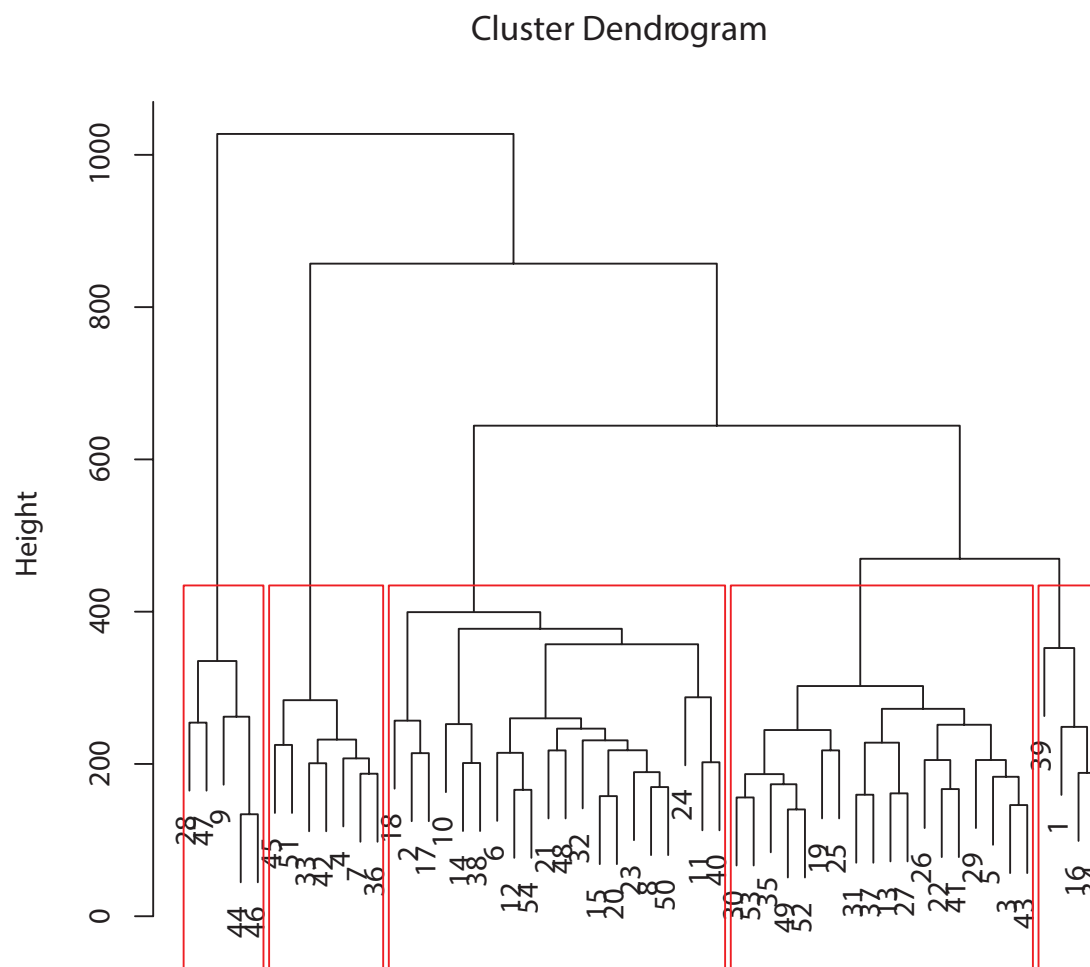


Figure 3: Dendrogram for the 2016 African data without grouping.

Table 3: Results of Hierarchical Clustering for the 54 Countries in 2016.

Country	IIAG Ranks	HC Groups	Country	IIAG Ranks	HC Groups
Botswana	2	1	Uganda	19	3
Cabo Verde	3	1	Zambia	13	3
Mauritius	1	1	Angola	45	4
Namibia	5	1	Burundi	43	4
Seychelles	4	1	Cameroon	38	4
South Africa	6	1	Chad	48	4
Tunisia	7	1	Comoros	26	4
Algeria	20	2	Congo	42	4
Egypt	24	2	D. R. of Congo	46	4
Morocco	14	2	Djibouti	37	4
Rwanda	9	2	Equatorial Guinea	47	4
Benin	16	3	Eritrea	50	4
Burkina Faso	23	3	Gabon	32	4
Cote d'Ivoire	21	3	Gambia	35	4
Ethiopia	31	3	Guinea	41	4
Ghana	8	3	Guinea-Bissau	44	4
Kenya	12	3	Mauritania	40	4
Lesotho	15	3	Nigeria	36	4
Liberia	28	3	Sao Tome & Principe	11	4
Madagascar	34	3	Swaziland	29	4
Malawi	17	3	Togo	33	4
Mali	25	3	Zimbabwe	39	4
Mozambique	22	3	Cent. African Republic	52	5
Niger	27	3	Libya	51	5
Senegal	10	3	Somalia	54	5
Sierra Leone	30	3	South Sudan	53	5
Tanzania	18	3	Sudan	49	5

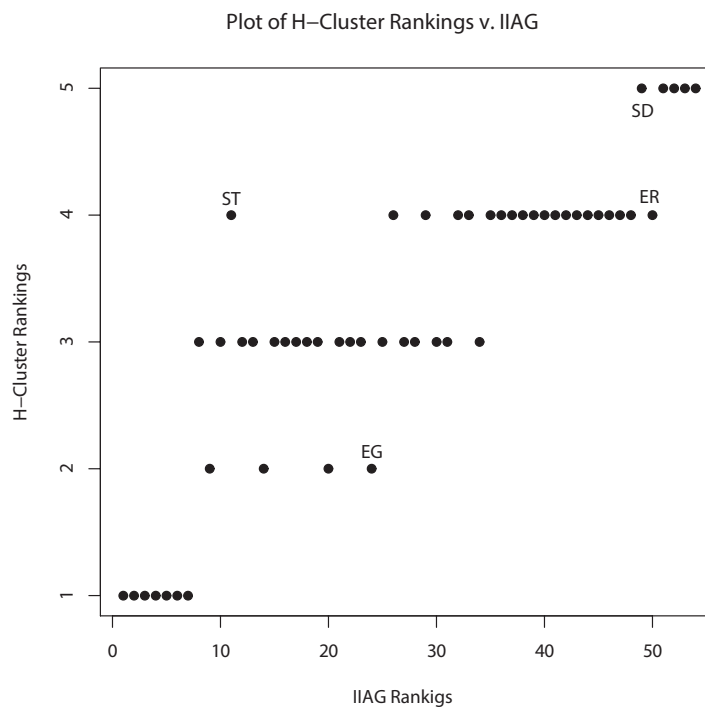


Figure 4: Plot of Hierarchical Cluster Rankings v. IIAG Rankings.

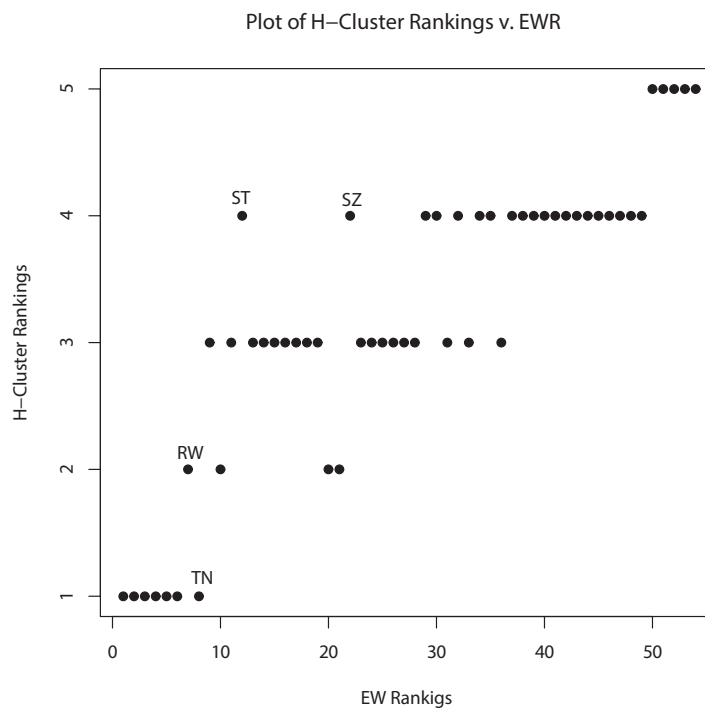


Figure 5: Plot of H-Cluster Rankings v. Equal Weight Rankings.

4 Summary and Conclusions

The contributions of this paper can be summarized as follows:

1. As an alternative methodology to computing composite indices, we propose the use of clustering methods to group objects or countries into K groups such that the countries in one group are similar to each other but different from the countries in the other groups. The use of clustering avoids two issues with composite indices, that is, the need to group variables/indicators by experts and the need to compute the standard errors. Countries in the same group are taken to be similar (having approximately the same rank) and countries that are classified into different groups are interpreted as having different ranks.
2. Different ways of ranking the countries are compared: The IIAG rankings, the Equal Weight rankings, the Principal Components Index rankings, and the Hierarchical Cluster rankings. These methods provide different rankings. Experts in the relevant domains can use these results and decide on which method gives appropriate rankings based on their theoretical understanding of the domain of interest (e.g., African governance for the African governance data).

References

1. Abd ElAziz, T. E. R. (2017), “Generalizing Almost First Order Dominance to the Multidimensional Case with Application on Poverty in Egypt,” Ph.D. Thesis, Faculty of Economics and Political Science, Cairo University.
2. Abo Elazm, H. F., (2017), “Measuring Happiness Among Youth (18–29 Years) in Egypt in 2009,” M. Sc. Thesis, Faculty of Economics and Political Science, Cairo University.
3. Bandura R. (2008), “A Survey of Composite Indices Measuring Country Performance,” United Nations Development Program, Office of Development Studies.
4. Billor, N., Hadi, A. S., and Velleman, P. F. (2000), “Blocked Adaptive, Computationally-Efficient outlier Numerator,” *Computational statistics and data analysis*, 34, 279–298.

5. Cronbach, L. J. (1951), “Coefficient alpha and the internal structure of tests,” *Psychometrika*, 16, 297–333.
6. Decanco, K., and Lugo, A. M. (2013) “Weights in multidimensional indices of well-being (An overview),” *Econometric Reviews*, 32, 7–34.
7. Dobbie, M. J., and Dail, D. (2013), “Robustness and sensitivity of weighting and aggregation in constructing composite indices,” *Ecological Indicators*, 29, 270–277.
8. 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.
9. Efron, B. and Tibshirani, R. (1993), “An Introduction to Bootstrap,” Chapman and Hall: New York.
10. Johnson, R. A., and Wichern, D. W. (1992), “Applied Multivariate Statistical Analysis,” New Jersey: Prentice Hall.
11. Jolliffe, I. T., (2002), “Principle Components Analysis,” Springer: New York.
12. Organization for Economic Co-Operation and Development (OECD), (2008), “Handbook on Constructing Composite Indicators - Methodology and User Guide,” OECD.
13. Rabea, S. A. R. (2017), “Measuring Child Wellbeing in Egypt Using Micro-Data,” M.Sc. Thesis, Faculty of Economics and Political Science, Cairo University.
14. Rocke, D.M., and Woodruff, D. L., (1996), “Identification of Outliers in Multivariate Data,” *Journal of the American Statistical Association*, 91, 1047–1061.
15. Rousseeuw, P. J. (1987), “Silhouettes: A Graphical Aid to the Interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, 20, 53–65.

16. Rousseuw, P. J., and Van Driessse, K., (1999), “A Fast Algorithm for Minimum Covariance Determinant Estimator,” *Technometrics*, 41, 212–223.

List of Sites:

17. Al Maktoum Foundation <http://WWW.mbrf.ae>, Last access 5/12/2017.
18. http://old.usc.ac.ir/IPPWebV1C035/TemplateFileFolder/10-8-2013/OriginalFolder/0a6e29c4-1fa6-4496-a362-000f0851d4fc_indices_2008_bandura.pdf.
Last access 5/12/2017.
19. IIAG: <http://www.moibrahimfoundation.org>, Last access 5/12/2017.

Review: Conditional and Unconditional AR (1) Model

Sayed M. El-Sayed¹, Ahmed A. El-Sheikh²,
Mohamed K. A. Issa³ and Hadia F.M.A. Azmy⁴

Abstract

In this paper, a literature of conditional and unconditional autoregressive model AR (1) will be discussed and explained which estimator is the best for application.

Keywords: First order autoregressive, Ordinary least squares, Maximum likelihood, Weighted symmetric.

1. Introduction

There are some similar aspects between autoregressive model and linear regression model, where y_t is the dependent variable, while $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ are the independent variables, and ε_t is considered to be the error term in autoregressive model. On the other hand, there is a difference between the two models; where the independent variables in auto-regressive model are random and correlated and there is another difference that the initial value (y_0) in AR model, this problem summarized in the first term of error (ε_1), because the value of ε_1 depends on the initial value y_0 and it cannot be estimated from the available observations $y_1, y_2, \dots, y_n$, but if $n \rightarrow \infty$ the effect of ε_1 will be negligible and so it will tend to zero. So there are different methods of estimation applied with the autoregressive models when y_0 is fixed, this method of estimation is called conditional estimation and when y_0 is stochastic

¹Prof of Applied Statistics, ISSR, Cairo University.

²Prof of Applied Statistics, ISSR, Cairo University.

³Assist. Prof of Applied Statistics, the Higher Institute of Cooperative and Managerial Studies.

⁴PhD. Student, ISSR, Cairo University.

and $N\left(0, \frac{\sigma^2}{1-\rho^2}\right)$, this method of estimation is called unconditional estimation [Sharawy (2005)].

The review is divided into two parts: the first part introduces a literature review of conditional AR (1) model, the second part introduces a literature review of unconditional AR (1) model.

2. Conditional AR (1) Model

Dickey and Fuller (1979) introduced the first order autoregressive model AR(1) which is defined as

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (1)$$

Where $t = 1, 2, \dots, n$, y_0 is fixed constant, and $\varepsilon_t \sim iid N(0, \sigma^2)$. They investigated the properties of the estimator ρ using the ordinary least squares method (OLS) under the assumption that $\rho = \pm 1$. They derived the limit distribution of the estimator of ρ . The results of Monte Carlo simulation indicated that tests based on the estimated ρ were more powerful for tests against stationarity than the $\hat{t}$ statistics.

Fuller and Hasza (1979) investigated an estimator for AR (1) model

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t \quad (2)$$

Where $t = 1, 2, \dots, n$, y_0 is fixed constant, and $\varepsilon_t \sim iid N(0, \sigma^2)$. They studied alternative predictors for stationary and non-stationary models using the Monte Carlo method. They showed that the OLS predictor is unbiased for symmetric error distributions for AR (1) model.

Fujikoshi and Ochi (1984) introduced the AR (1) model which is defined in (1) Where $t = \dots, -1, 0, 1, \dots$ and ρ is an unknown parameter, $|\rho| < 1$ and $\varepsilon_t \sim iid N(0, \sigma^2)$. They derived an asymptotic expansion of the distribution of

Maximum likelihood (ML) estimator based on T observations for the AR (1) normal model up to the term of order T^{-1} .

Dickey, et al. (1984) proposed the simple weighted symmetric (SWS) estimator for AR (1) model without constant as defined in (1). Where $t=1,2,\dots,n$ and y_0 is fixed constant, $\varepsilon_t \sim iid N(0, \sigma^2)$. The SWS estimator of ρ is given by

$$\hat{\rho}_{SWS} = - \left[0.5(y_1^2 + y_n^2) + \sum_{t=2}^{n-1} y_t^2 \right]^{-1} \sum_{t=2}^n y_t y_{t-1}$$

Breton and Pham (1989) introduced an estimator for the AR (1) model without constant as defined in (1). Where $t \geq 1$, $y_0 = 0$, ρ is a real parameter and $\varepsilon_t \sim iid N(0, \sigma^2)$. They obtained the exact formula for the bias of estimator and derived the asymptotic bias using OLS and ML.

Park and Fuller (1995) introduced a weighted symmetric (WS) estimator for AR(1) model without constant as defined in (1). Where $t = 2, 3, \dots, n$, $\varepsilon_t \sim iid N(0, \sigma^2)$ and y_0 is fixed constant. Under the stationary condition, that is, $|\rho| = 1$ could have a backward representation of the autoregressive time series. That is,

$$y_t = \rho y_{t+1} + \varepsilon_t \quad t = 2, 3, \dots, n \quad (3)$$

The WS estimator of ρ denoted by $\hat{\rho}_{WS}$, is derived by minimizing the following function

$$Q_{WS}(\rho) = \sum_{t=2}^n w_t (y_t - \rho y_{t-1})^2 + \sum_{t=1}^{n-1} (1 - w_{t+1}) (y_t - \rho y_{t+1})^2$$

The WS estimator for AR (1) as follows:

$$\hat{\rho}_{WS} = \frac{\sum_{t=2}^n y_t y_{t-1}}{\sum_{t=2}^{n-1} y_t^2 + n^{-1} \sum_{t=1}^n y_t^2}$$

Shenton and Vinod (1995) conducted closed forms for asymptotic bias and variance for AR (1) model without constant as defined in (1) with unit roots and with

zero initial value. They provided the closed- form for the asymptotic bias and variance of ML estimator as follows;

$$\hat{\rho} = \frac{\sum_1^n y_t y_{t-1}}{\sum_1^{n-1} y_t^2}$$

Where $\varepsilon_t \sim iid N(0, \sigma^2)$ and $|\rho| = 1$. Numerical accuracy up to six significant digits was achieved for sample sizes when $n > 100$.

Ali (2002) investigated the finite sample distribution of the OLS estimator of the autoregressive parameter in AR (1) model without constant as defined in (1), obtained a uniform asymptotic expansion for the distribution applicable to both stationary and non-stationary cases, used three cases of initial values y_0 , where $y_0 = 0$ is fixed, $y_0 \neq 0$ is constant and y_0 is stochastic.

Provost and Sanjel (2005) investigated the estimator of the coefficient of AR (1) model as defined in (1). Where $t = 0, \pm 1, \pm 2, \dots, \pm n$, $\varepsilon_t \sim iid N(0, \sigma^2)$ and $|\rho| = 1$. They compared estimator of the parameter of AR (1) model using the OLS, modified least squares (MLS), the Yule-Walker, and Burg's estimators. They showed that the MLS estimator is the least biased and the OLS and Burg's estimators have similar means and variances for any $n > 10$.

Youssef (2008) introduced an estimator for the AR (1) model with unit root as defined in (2). Where $t = 2, 3, \dots, n$, $\varepsilon_t \sim iid N(0, \sigma^2)$. He compared the OLS, MLS, WS, modified weighted symmetric (MWS), and ML estimators for AR (1) in the case of unit root and he studied it using Monte Carlo method, to see how well the estimators and the predictors performed on different samples sizes. The Monte Carlo simulation results suggested that the MLS estimator performed well, in the sense of less biased and MSE, in small and large sample size. Therefore, he recommend to use MLS to estimate the unit roots.

Also Youssef, et al (2012) introduced a comparison between some methods of estimation, such as (OLS), (MLS), (WS), (MWS) and unconditional maximum likelihood (UML) for AR (1) model as defined in (2). Where $t = 2, 3, \dots, n$,

$\varepsilon_t \sim iid N(0, \sigma^2)$. Also, the properties of OLS estimator were discussed. A comparative study between these methods was considered in the case of unit root and explosive time series, using Monte Carlo simulation. They concluded that the estimator of MLS method was the best one among other estimators for small and medium sample sizes, while UML was the best one for large sample size.

El-Sayed et al (2015) proved the linearization and the unbiasedness of the estimated parameter of AR (1) model without constant as defined in (1) and (3) using WS method, where $t = 2, 3, \dots, n$, $\varepsilon_t \sim iid N(0, \sigma^2)$, and $|\rho| = 1$.

3. Unconditional AR(1) Model

Box and Jenkins (1976) investigated ML estimator of the parameter for AR(1) model as defined in (1). They showed that there is no direct analytical formula in terms of series data when derive the logarithmic of unconditional maximum likelihood function (4), so the following function has to be solving by one of numerical methods.

$$\ln L = -\frac{n}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \ln(1 - \rho^2) - \frac{1}{2\sigma^2} \left[y_1^2(1 - \rho^2) + \sum_{t=2}^n (y_t - \rho y_{t-1})^2 \right] \quad (4)$$

They showed that the fundamental reason in the absence of such formula due to the presence of quantity $\ln(1 - \rho^2)$ in the logarithmic function, so assumed that the effect of $\ln(1 - \rho^2)$ will be negligible and so it will tend to zero if $n \rightarrow \infty$. Then the logarithm of likelihood function can be written in approximate way as the follow.

$$\ln(L) \approx -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left[y_1^2(1 - \rho^2) + \sum_{t=2}^n (y_t - \rho y_{t-1})^2 \right]$$

And they obtain the UML estimator as follows;

$$\hat{\rho}_{UML} = \frac{\sum_{t=2}^n y_t y_{t-1}}{\sum_{t=2}^n y_{t-1}^2}$$

They concluded that the estimator $\hat{\rho}_{UML}$ is the same as the unconditional ordinary least squares (UOLS) estimator $\hat{\rho}_{UOLS}$.

Hasza (1980) derived approximate estimator for AR (1) for $\rho \in (-\infty, \infty)$ and showed that there is a root in each of interval $(-\infty, -1)$, $(-1, 1)$, and $(1, \infty)$ by using ML. If y_t is stationary, then the estimator for AR(1) for ρ is a root of the following cubic equation

$$f(\rho) = a_3\rho^3 + a_2\rho^2 + a_1\rho + a_0$$

Where

$$a_3 = \left[\frac{n-1}{n} \right] \sum_{t=2}^{n-1} y_t^2, \quad a_2 = - \left[\frac{n-2}{n} \right] \sum_{t=2}^n y_t y_{t-1}, \quad a_1 = \sum_{t=2}^{n-1} y_t^2 - \frac{1}{n} \sum_{t=1}^n y_t^2, \quad a_0 = \sum_{t=2}^n y_t y_{t-1}$$

Shin and Fuller (1998) proposed a Likelihood ratio test based on UML estimation for autoregressive moving average (ARMA) that may contained an AR unit root, and this was considered as an extension of the work of Hasza (1980). They showed that the limiting distribution of normalized ML estimator of the unit root was the same as the estimator for AR (1) model.

El-Sayed, et al. (2017) investigated a closed form of the bias of the parameter using UML of AR(1) model as defined in (1) Where $t = 1, 2, 3, \dots, n$, $\varepsilon_t \sim iidN(0, \sigma^2)$, assumed that y_0 is $N(0, \frac{\sigma^2}{1-\rho^2})$ under the stationarity condition $|\rho| < 1$, which is considered as an extension of Box and Jenkins (1976). Moreover, the consistency of the estimated parameter has been proved.

4. Conclusion

Based on the previous review, the following points can be presented, as;

- For AR (1) model with unit root problem, the MLS method was the best one among other estimators for small and medium sample sizes, while UML method was the best one for large sample size [Youssef (2008)] and [Youssef, et al (2012)].
- For AR (1) model with explosive problem, the UML method which considered the only method of estimation that can be used in this case [Youssef, et al (2012)].
- For stationary AR (1) model a closed form of the bias using UML as the method estimation has been derived [El-Sayed, et al. (2017)].

References

1. Ali, M.M. (2002). Distribution of The Least Squares Estimation in a First Order Autoregressive Model. *Econometric Reviews*, 21(1), 89-119.
2. Box, G. E. P. and Jenkins, G. M. (1976). *Time series analysis forecasting and control*. San Francisco, Holden Day
3. Breton, A. L. and Pham, D. T. (1989). On the bias of the least squares estimator for the first order autoregressive process. *Ann. Inst. Statist. Math.* 41(3), 555-563.
4. Dickey, D.A. and Fuller, W.A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *JASA*, 74(366), 427-431.
5. Dickey, D. A., Haszaand, D. P. and Fuller, W. A. (1984). Testing for unit roots in seasonal time series, *JASA*, 79, 355-367.
6. El-Sayed, S.M., El-Sheikh, A. A. and Issa, M. K. A. (2015). Weighted Symmetric Estimators of Autoregressive Models. *Far East Journal of Advances and Applications in Statistics*, 47(2), 145-152.
7. El-Sayed, S.M., El-Sheikh, A. A., Issa, M. K. A. and Azmy, H.F.M.A (2017). A Closed Form of Biased AR (1) Model. *Far East Journal of Advances and Applications in Statistics*, 50(3), 191-199.

8. Fujikoshi, Y. and Ochi, Y. (1984). Asymptotic Properties of the Maximum Likelihood Estimate in the First Order Autoregressive Process. *Ann. Inst. Statist. Math. P.A*, 119-128.
9. Fuller, W.A. and Hasza, D.P. (1979). Predictors for the First-Order Autoregressive Process. *Journal of Econometrics*, 13, 139-157.
10. Hasza, D.P. (1980). A Note on Maximum Likelihood Estimation for the First Order Autoregressive Process. *Communication in Statistics, Theory and Methods*, 9 (13), 1411-1415.
11. Park, H. and Fuller, W.A. (1995). Alternative Estimators and Unit Root Tests for the Autoregressive Process. *Journal of Time Series Analysis*. 16, 415-429.
12. Provost, S.B. and Sanjel, S. (2005). Inference About the First Order Autoregressive Coefficient. *Communication in Statistics-Theory and Method*, 34(5), 1183-1201.
13. Sharawy, S.M. (2005). *An Introduction of Modern Time Series Analysis*. Scientific Research Centre, Faculty of Science, King Abd-Elaziz University, Jeddah, Kingdom of Saudi Arabia.
14. Shenton. L.R. and Vinod, H.D. (1995). Closed Forms for Asymptotic Bias and Variance in Autoregressive Models with Unit Roots, *Journal of Computational and Applied Mathematics*. 61, 231-243.
15. Shin, D.W. and Fuller, W.A. (1998). A unit Root Tests Based on Unconditional Maximum Likelihood Estimation for the Autoregressive and Moving Average. *Journal of Time Series Analysis*, 19(5), 591-599.
16. Youssef, A. H. (2008). A New Estimator for the Unit Root. *Journal of Statistical Computation and Simulation*, 78, 515 – 526.
17. Youssef, A. H., El-Sheikh, A. A and Issa, M. K. A. (2012). A Comparison of Estimators for the First Order Autoregressive Process with a Unit Root. *Far East Journal of Theoretical Statistics*, 41(2), 133-147.

Estimation of Ranked Set Sampling Methods : Comparative Study

Nasr I. Rashwan¹ and Nuran M. Al-Mawan^{2,3}

Abstract :

In environmental monitoring and assessment, the main focus is to achieve observational economy and to collect data with unbiased, efficient and cost effective sampling methods. Ranked set sampling (RSS) is one traditional method that is mostly used for accomplishing observational economy. In this article, quartile double ranked set sampling (QDRSS) is compared to different ranked set sampling methods which are useful to estimate the population mean based on simple random sampling (SRS) method. When the underlying distributions are symmetric and asymmetric, it turns out that quartile double RSS (QDRSS) produce unbiased estimator is more efficient than SRS, paired RSS (PPRSS), extreme RSS (ERSS), median RSS (MRSS), double RSS (DRSS), quartile RSS (QRSS), percentile RSS (PRSS), paired double RSS (PPDRSS) and RSS based on the sample size. It is shown that, the variance of the mean estimator under QDRSS is always less than the variance of the mean estimator based on SRS and the other methods.

Keywords:

Ranked set sampling; Paired ranked set sampling; Extreme ranked set sampling; Median ranked set sampling; Double ranked set sampling; Quartile ranked set sampling; Percentile ranked set sampling; Quartile double ranked set sampling; Paired double ranked set sampling; Relative precision.

1. Introduction

There are many sampling methods that can be used in surveys of natural resources in agriculture, biology, environmental management, forestry, etc. The main objective of whatever sampling method is used is to obtain precise estimates of population parameters with minimum cost and expenditure. One method is ranked set sampling (RSS). RSS becomes an efficient alternative to simple random sampling (SRS) when taking exact measurements of selected units is very costly whereas ranking a small set of selected units is cheap. Ranking may be visually with respect to the study variable or by any inexpensive method. McIntyre^[10] was the first to suggest the RSS method for estimation of pasture and forage yields. McIntyre's goal was to maintain the unbiased of SRS while effectively incorporating into the estimates the outside information provided by visual inspection. Takahasi and Wakimoto^[21] developed the theory of RSS procedure under the assumption of perfect ranking. Dell and Clutter^[5] showed that the mean estimator under imperfect RSS remains an unbiased estimator

Department of Statistics & Mathematics , Faculty of Commerce , Tanta University , Tanta , Egypt
Department of Applied Statistics and Econometric , Institute of Statistical Studies and Research ,Cairo University ,
Cairo , Egypt

Department of Basic Science , Faculty of Engineering ,Modern Academy , Cairo , Egypt

of population mean. Patil et al. ^[16] compared the mean estimator based on perfect and imperfect rankings with the SRS based regression estimator of population mean. Muttalak ^[11] and Muttalak ^[12] introduced paired ranked set sampling (PPRSS) and median ranked set sampling (MRSS) schemes for estimation of population mean, respectively. PPRS can be used when it is difficult to select a large number of units from the population of interest. Samawi et al. ^[19] suggested using extreme ranked set sampling (ERSS) to estimate the population mean if the underlying distribution is symmetric and it is more efficient than the SRS estimator. Muttalak ^[13] extended the work of Patil et al. ^[16] and showed that the mean estimator under MRSS is more efficient than the mean estimators with SRS and RSS. Al-Saleh and Al-Kadiri^[1] suggested double ranked set sampling method (DRSS) for estimating the population mean. They showed that the sample mean based on DRSS is more efficient than the sample mean with RSS. Muttalak ^[14,15] proposed quartile ranked set sampling (QRSS) and percentile ranked set sampling (PRSS) for estimating the population mean. Al-Saleh and Al-Omari ^[2] introduced quartile double ranked set sampling (QDRSS) procedure for estimating the population mean. Haq et al. ^[7] suggested paired double ranked set sampling (PPDRSS) for estimating the population mean. The main advantage of using PPDRSS over DRSS is that it requires less number of identified units as compared with DRSS. This helps in reducing the time and cost that is involved in the ranking process.

In this paper, quartile double ranked set sampling (QDRSS) method is compared with the other ranked set sampling methods. It is shown that for the probability distributions considered in this study, QDRSS unbiased estimator has less variance than all the other methods. Computer simulation results are given to compare the efficiency for the estimators based on SRS with its counterparts RSS, PPRS, ERSS, MRSS, QRSS, PRSS, DRSS, QDRSS and PPDRSS.

2. Ranked set sampling methods

This section is concerned with methods of ranked set sampling given a definition of each method and defined the mean and the variance of each case in the methods.

Let $X_1, X_2, \dots, X_n$ be a random sample with probability density function $f(x)$ with mean μ and variance σ^2 . Let $X_{1(1:n)}, X_{2(1:n)}, \dots, X_{n(1:n)}$; $X_{1(2:n)}, X_{2(2:n)}, \dots, X_{n(2:n)}$; ... ; $X_{1(n:n)}, X_{2(n:n)}, \dots, X_{n(n:n)}$ be independent random variables all with the same cumulative distribution function $F(x)$. Let $X_{i(i:n)}$ denotes the i^{th} order statistic from the i^{th} sample of size n ($i=1, 2, \dots, n$). The cycle may be repeated m times to get mn units. Let $X_{(i:nj)}$ denotes the j^{th} cycle of size m ($j=1, 2, \dots, m$). To understand the calculation in the study assume the cycle is repeated once ; i.e. $m=1$. The variance or the mean square error of the sample means for the RSS, PPRS, ERSS, MRSS, QRSS, PRSS, DRSS, QDRSS and PPDRSS were calculated using the moments of the order statistics for the distributions considered in the study (see Harter and Balakrishnan ^[8] and

Balakrishnan and Chen ^[3]). The relative efficiency (RE) of estimating the population mean using any of the RSS based methods with respect to the usual estimator using SRS is defined as ^[14]

$$RE(\bar{X}_{SRS}, \bar{X}_{RSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{RSS})}$$

Here $\bar{X}_{SRS}$ and $\bar{X}_{RSS}$ are both unbiased estimators. In case of biased estimator, the variance is replaced by the *MSE* of the estimator where ^[17]

$$MSE(\bar{X}_{RSS}) = Var(\bar{X}_{RSS}) + [E(\bar{X}_{RSS}) - \mu]^2.$$

2.1. Ranked set sampling

The original idea of ranked set sampling (RSS) was introduced by McIntyre's ^[10]. Ranked set sampling is a methodology which can improve the efficiency of techniques such as estimation and confidence intervals without increasing the number of substantial observations. It uses additional information from the population to provide more structure to the data collection process and decreases the likelihood of an unrepresentative sample. In addition, it is designed to minimize the number of measured observations required to achieve the desired precision in making inferences. The RSS scheme can be described as follows : In order to draw a sample of size n , we identify n^2 units from the target population. The n units in each sample are ranked visually or by any inexpensive method with respect to the variable of interest. From the first set of n units, the smallest ranked unit is measured. From the second set of n units, the second smallest ranked unit is measured. The process is continued until from the n th set of n units the largest ranked unit is measured. The estimator of the population mean for RSS of size n is given by

$$\bar{X}_{RSS} = \frac{1}{n} \sum_{i=1}^n X_i,$$

with variance

$$Var(\bar{X}_{RSS}) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2_{(i:n)}, \text{ where } \sigma^2 = E(X_{(i:n)} - \mu_{(i:n)})^2 = \sigma^2 - \sum_{i=1}^n (\mu_{(i:n)} - \mu)^2.$$

The RE for the symmetric and asymmetric distributions are

$$RE(\bar{X}_{SRS}, \bar{X}_{RSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{RSS})}.$$

2.2. Paired ranked set sampling

Muttlak ^[11] introduced paired ranked set sampling (PPRSS) procedure for estimating the population mean. PPRSS procedure is explained as follows: for even sample size n , identify n^2 sets each of size n from the target population. Each sample is ranked in itself as in ranked set sampling design. Select the smallest and largest ranked units from the first set. Similarly, select the second smallest and second largest ranked units from the second set. The procedure continues, then $\frac{n}{2}$ th and $\frac{n+2}{2}$ th ranked units are selected from the last set. Similarly, for odd sample size n , identify $\frac{n+1}{2}$ sets each of size n from the target population. Each sample is ranked in itself as in ranked set sampling design. Select the smallest and largest ranked units

from the first set. Similarly, select the second smallest and second largest ranked units from the second set. The procedure continues, then the $\frac{n+1}{2}$ th ranked units are selected from the last set. The sample means under PPRSS for even and odd n are given below:^[11]

$$\bar{X}_{PPRSSE} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n}{2}} X_{i(n)} + \sum_{i=1}^{\frac{n}{2}} X_{i(n-i+1:n)} \right)$$

$$\text{and } \bar{X}_{PPRSSO} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n+1}{2}} X_{i(n)} + \sum_{i=1}^{\frac{n-1}{2}} X_{i(n-i+1:n)} \right),$$

Both $\bar{X}_{PPRSSE}$ and $\bar{X}_{PPRSSO}$ are unbiased estimators. The variances of $Var(\bar{X}_{PPRSSE})$ and $Var(\bar{X}_{PPRSSO})$ respectively, are

$$Var(\bar{X}_{PPRSSE}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma^2_{(i:n)} + \sum_{i=1}^{\frac{n}{2}} \sigma^2_{(n-i+1:n)} \right),$$

$$Var(\bar{X}_{PPRSSO}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n+1}{2}} \sigma^2_{(i:n)} + \sum_{i=1}^{\frac{n-1}{2}} \sigma^2_{(n-i+1:n)} \right).$$

The RE for the symmetric distribution is

$$RE(\bar{X}_{SRS}, \bar{X}_{PPRSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{PPRSS})},$$

$$\text{and } RE(\bar{X}_{SRS}, \bar{X}_{PPRSS}) = \frac{Var(\bar{X}_{SRS})}{MSE(\bar{X}_{PPRSS})} \text{ for asymmetric distribution.}$$

2.3. Extreme ranked set sampling

Samawi et al. ^[19] used extreme ranked set sample (ERSS) in case of even sample size with is easier to use than the usual RSS procedure to estimate the population mean. ERSS procedure is explained as follows: for even sample size n , identify n^2 units from the target population. Each sample is ranked in itself as in ranked set sampling design. Select the smallest ranked unit $X_{(1)}$ from first $\frac{n}{2}$ sets and the largest ranked unit $X_{(n)}$ from the other $\frac{n}{2}$ sets. Similarly, for odd sample size n , select the smallest ranked unit $X_{(1)}$ from first $\frac{n-1}{2}$ sets. Select the largest ranked unit $X_{(n)}$ from second $\frac{n-1}{2}$ sets and the median with rank $\frac{n+1}{2}$ from the remaining set. The sample means under ERSS depending on even (E) and odd (O) sample sizes are, respectively, given by ^[19]

$$\bar{X}_{ERSSE} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n}{2}} X_{i(1:n)} + \sum_{i=\frac{n}{2}+1}^n X_{i(n:n)} \right)$$

$$\text{and } \bar{X}_{ERSO} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n-1}{2}} X_{i(1:n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} X_{i(n:n)} + X_{i(\frac{n+1}{2}:n)} \right),$$

Both $\bar{X}_{ERSSE}$ and $\bar{X}_{ERSO}$ are unbiased estimators. The variances of $Var(\bar{X}_{ERSSE})$ and $Var(\bar{X}_{ERSO})$ respectively, are

$$Var(\bar{X}_{ERSSE}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma^2_{(1:n)} + \sum_{i=\frac{n}{2}+1}^n \sigma^2_{(n:n)} \right)$$

$$Var(\bar{X}_{ERSO}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n-1}{2}} \sigma^2_{(1:n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} \sigma^2_{(n:n)} + \sigma^2_{(\frac{n+1}{2}:n)} \right), \quad (2.1)$$

The RE for the symmetric distribution is

$$RE(\bar{X}_{SRS}, \bar{X}_{ERSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{ERSS})},$$

$$\text{and } RE(\bar{X}_{SRS}, \bar{X}_{ERSS}) = \frac{Var(\bar{X}_{SRS})}{MSE(\bar{X}_{ERSS})} \text{ for asymmetric distribution.}$$

2.4. Median ranked set sampling

Muttalak ^[12] introduced Median ranked set sampling (MRSS) procedure for estimating the population mean. MRSS procedure is explained as follows: for even sample size n , identify n^2 units from the target population. Each sample is ranked in itself as in ranked set sampling design. Select the $\frac{n}{2}th$ smallest rank from the first $\frac{n}{2}$ sets and select the $\frac{n+2}{2}th$ smallest rank from the other $\frac{n}{2}$ sets. Similarly, for odd sample size n , select the $\frac{n+1}{2}th$ smallest rank from all sets. The sample means under MRSS for even and odd n are given below:^[7]

$$\bar{X}_{MRSSe} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n}{2}} X_{i(\frac{n}{2}:n)} + \sum_{i=\frac{n}{2}+1}^n X_{i(\frac{n+2}{2}:n)} \right)$$

$$\text{and } \bar{X}_{MRSSo} = \frac{1}{n} \left(\sum_{i=1}^n X_{i(\frac{n+1}{2}:n)} \right),$$

Both $\bar{X}_{MRSSe}$ and $\bar{X}_{MRSSo}$ are unbiased estimators. The variances of $Var(\bar{X}_{MRSSe})$ and $Var(\bar{X}_{MRSSo})$ respectively, are

$$Var(\bar{X}_{MRSSe}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma^2_{(\frac{n}{2}:n)} + \sum_{i=\frac{n}{2}+1}^n \sigma^2_{(\frac{n+2}{2}:n)} \right),$$

$$Var(\bar{X}_{MRSSo}) = \frac{1}{n^2} \left(\sum_{i=1}^n \sigma^2_{(\frac{n+1}{2}:n)} \right) \quad (2.2)$$

The RE for the symmetric distribution is

$$RE(\bar{X}_{SRS}, \bar{X}_{MRSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{MRSS})},$$

$$\text{and } RE(\bar{X}_{SRS}, \bar{X}_{MRSS}) = \frac{Var(\bar{X}_{SRS})}{MSE(\bar{X}_{MRSS})} \text{ for asymmetric distribution.}$$

2.5. Double ranked set sampling

Al-Saleh and Al-Kadiri ^[1] introduced double ranked set sampling (DRSS) procedure for estimating the population mean. In order to select a double ranked set sample of size n , the experimenter needs to identify n^3 units from the target population. This may be difficult when an epidemic breaks out in some area or in queuing problems when data arrive in batches of varying sizes. Moreover, there may be a shortage of experimental units or ranking is difficult, time consuming and costly, see Samawi ^[18] and Haq et al. ^[6]. In all such situations, the balanced DRSS cannot be conducted with full confidence or it is costly. The DRSS scheme can be described as follows: In order to draw a sample of size n , we identify n^3 units from the target population. Randomly allocate these units to n set each of size n^2 . Apply RSS procedure to n sets in order to obtain n ranked set samples each of size n . Again apply the RSS procedure to these n ranked sets, each of size n , to obtain a double ranked set sample of size n . The estimator of the population mean for DRSS of size n is given by ^[7]

$$\bar{X}_{DRSS} = \frac{1}{n} \sum_{i=1}^n X_{i^*}^*,$$

with variance

$$Var(\bar{X}_{DRSS}) = \frac{1}{n^2} \sum_{i=1}^n \sigma^{*2}_{(i:n)}, \text{ where } \sigma^{*2}_{(i:n)} = E(X_{(i:n)} - \mu_{(i:n)})^2.$$

The RE for the symmetric distribution is

$$RE(\bar{X}_{SRS}, \bar{X}_{DRSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{DRSS})},$$

$$\text{and } RE(\bar{X}_{SRS}, \bar{X}_{DRSS}) = \frac{Var(\bar{X}_{SRS})}{MSE(\bar{X}_{DRSS})} \text{ for asymmetric distribution.}$$

2.6. Quartile ranked set sampling

Muttlak ^[14] introduced quartile ranked set sampling (QRSS) procedure for estimating the population mean. QRSS procedure is explained as follows: for even sample size n , identify n^2 units from the target population. Each sample is ranked in itself as in ranked set sampling design. Select the $(q_1(n+1))$ th smallest ranked unit from the first $\frac{n}{2}$ sets and select the $(q_3(n+1))$ th smallest ranked unit from the other $\frac{n}{2}$ sets where $q_1 = 0.25$ and $q_3 = 0.75$. Similarly, for odd sample size n , select the $(q_1(n+1))$ th smallest ranked unit from first $\frac{n-1}{2}$ sets. Select the $(q_3(n+1))$ th smallest ranked unit from second $\frac{n-1}{2}$ sets and the median with rank $\frac{n+1}{2}$ from the remaining set. The sample means under QRSS for even and odd n are given below:^[14]

$$\bar{X}_{QRSSE} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n}{2}} X_{i(q_1(n+1):n)} + \sum_{i=\frac{n}{2}+1}^n X_{i(q_3(n+1):n)} \right)$$

$$\text{and } \bar{X}_{QRSSO} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n-1}{2}} X_{i(q_1(n+1):n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} X_{i(q_3(n+1):n)} + X_{i(\frac{n+1}{2}:n)} \right),$$

Both $\bar{X}_{QRSSE}$ and $\bar{X}_{QRSSO}$ are unbiased estimators. The variances of $Var(\bar{X}_{QRSSE})$ and $Var(\bar{X}_{QRSSO})$ respectively, are

$$Var(\bar{X}_{QRSSE}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma^2_{(q_1(n+1):n)} + \sum_{i=\frac{n}{2}+1}^n \sigma^2_{(q_3(n+1):n)} \right) \quad (2.3)$$

$$Var(\bar{X}_{QRSSO}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n-1}{2}} \sigma^2_{(q_1(n+1):n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} \sigma^2_{(q_3(n+1):n)} + \sigma^2_{(\frac{n+1}{2}:n)} \right), \quad (2.4)$$

The RE for the symmetric distribution is

$$RE(\bar{X}_{SRS}, \bar{X}_{QRSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{QRSS})},$$

$$\text{and } RE(\bar{X}_{SRS}, \bar{X}_{QRSS}) = \frac{Var(\bar{X}_{SRS})}{MSE(\bar{X}_{QRSS})} \text{ for asymmetric distribution.}$$

2.7. Percentile ranked set sampling

Muttlak ^[15] introduced percentile ranked set sampling (PRSS) procedure with different value of $0 \leq p \leq 1$ for estimating the population mean. PRSS procedure is explained as follows: for even sample size n , identify n^2 units from the target population. Each sample is ranked in itself as in ranked set sampling design. Select the $(p(n+1))$ th smallest ranked unit from the first $\frac{n}{2}$ sets and select the $(q(n+1))$ th smallest ranked unit from the other $\frac{n}{2}$ sets. Similarly, for odd sample size n , select the $(p(n+1))$ th smallest ranked unit from first $\frac{n-1}{2}$ sets. Select the $(q(n+1))$ th smallest ranked unit from second $\frac{n-1}{2}$ sets and the median with rank $\frac{n+1}{2}$ from the remaining set. The sample means under PRSS for even and odd n are given below:^[15]

$$\bar{X}_{PR SSE} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n}{2}} X_{i(p(n+1):n)} + \sum_{i=\frac{n}{2}+1}^n X_{i(q(n+1):n)} \right)$$

$$\text{and } \bar{X}_{PRSSO} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n-1}{2}} X_{i(p(n+1):n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} X_{i(q(n+1):n)} + X_{i(\frac{n+1}{2}:n)} \right),$$

Both $\bar{X}_{PR SSE}$ and $\bar{X}_{PRSSO}$ are unbiased estimators. The variances of $Var(\bar{X}_{PR SSE})$ and $Var(\bar{X}_{PRSSO})$ respectively, are

$$Var(\bar{X}_{PRSE}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma^2_{(p(n+1):n)} + \sum_{i=\frac{n}{2}+1}^n \sigma^2_{(q(n+1):n)} \right) \quad (2.5)$$

$$Var(\bar{X}_{PRSSO}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n-1}{2}} \sigma^2_{(p(n+1):n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} \sigma^2_{(q(n+1):n)} + \sigma^2_{(\frac{n+1}{2}:n)} \right), \quad (2.6)$$

The RE for the symmetric distribution is

$$RE(\bar{X}_{SRS}, \bar{X}_{PRSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{PRSS})},$$

and $RE(\bar{X}_{SRS}, \bar{X}_{PRSS}) = \frac{Var(\bar{X}_{SRS})}{MSE(\bar{X}_{PRSS})}$ for asymmetric distribution.

2.8. Quartile double ranked set sampling

Al-Saleh and Al- Omari ^[2] introduced quartile double ranked set sampling (QDRSS) procedure for estimating the population mean. The QDRSS scheme can be described as follows : In order to draw a sample of size n , we identify n^3 units from the target population. Randomly allocate these units to two sets each of size n^2 . Apply RSS procedure to n sets in order to obtain n ranked set samples each of size n . Apply the QRSS procedure to these n ranked sets, each of size n , to obtain a quartile double ranked set sample of size n . The sample means under QDRSS for even and odd n are given below:^[2]

$$\bar{X}_{QDRSSE} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n}{2}} X^*_{i(q_1(n+1):n)} + \sum_{i=\frac{n}{2}+1}^n X^*_{i(q_3(n+1):n)} \right)$$

$$\text{and } \bar{X}_{QDRSSO} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n-1}{2}} X^*_{i(q_1(n+1):n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} X^*_{i(q_3(n+1):n)} + X^*_{i(\frac{n+1}{2}:n)} \right),$$

Both $\bar{X}_{QDRSSE}$ and $\bar{X}_{QDRSSO}$ are unbiased estimators. The variances of $Var(\bar{X}_{QDRSSE})$ and $Var(\bar{X}_{QDRSSO})$ respectively, are

$$Var(\bar{X}_{QDRSSE}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma^{*2}_{(q_1(n+1):n)} + \sum_{i=\frac{n}{2}+1}^n \sigma^{*2}_{(q_3(n+1):n)} \right)$$

$$Var(\bar{X}_{QDRSSO}) = \frac{1}{n} \left(\sum_{i=1}^{\frac{n-1}{2}} \sigma^{*2}_{(q_1(n+1):n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} \sigma^{*2}_{(q_3(n+1):n)} + \sigma^2_{(\frac{n+1}{2}:n)} \right),$$

The RE for the symmetric distribution is

$$RE(\bar{X}_{SRS}, \bar{X}_{QDRSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{QDRSS})},$$

and $RE(\bar{X}_{SRS}, \bar{X}_{QDRSS}) = \frac{Var(\bar{X}_{SRS})}{MSE(\bar{X}_{QDRSS})}$ for asymmetric distribution.

2.9. Paired double ranked set sampling

Haq et al. ^[7] introduced paired double ranked set sampling (PPDRSS) procedure for estimating the population mean. This method can be used as an alternative to DRSS when it's difficult or time consuming or costly to identify n^3 units from the target population, especially when there is a shortage of experimental units. The main steps involved in selecting PPDRSS of size n are as follows: for even sample size n , identify $\frac{n^3}{2}$ units from the target population. Each sample is ranked in itself as in ranked set sampling design. Randomly allocate these units to $\frac{n}{2}$ set of size n^2 . Apply RSS procedure on each set in order to obtain $\frac{n}{2}$ ranked set sample of size n . Now, apply PPRSS procedure on $\frac{n}{2}$ set to get an even PPDRSS of size n . Similarly, for odd sample size n , identify $\frac{n^2(n+1)}{2}$ units from the target population. Randomly

allocate these units to $\frac{n+1}{2}$ sets each of size n^2 . Apply RSS procedure on each set in order to obtain $\frac{n+1}{2}$ ranked set samples each of size n . Now, apply PPRSS procedure on $\frac{n+1}{2}$ sets to get an odd PPDRSS of size n . The sample means under PPDRSS for even and odd n are given below:^[7]

$$\bar{X}_{PPDRSSE} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n}{2}} X_{i(i:n)}^* + \sum_{i=1}^{\frac{n}{2}} X_{i(n-i+1:n)}^* \right)$$

$$\text{and } \bar{X}_{PPDRSSO} = \frac{1}{n} \left(\sum_{i=1}^{\frac{n+1}{2}} X_{i(i:n)}^* + \sum_{i=1}^{\frac{n-1}{2}} X_{i(n-i+1:n)}^* \right),$$

Both $\bar{X}_{PPDRSSE}$ and $\bar{X}_{PPDRSSO}$ are unbiased estimators. The variances of $Var(\bar{X}_{PPDRSSE})$ and $Var(\bar{X}_{PPDRSSO})$ respectively, are

$$Var(\bar{X}_{PPDRSSE}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma_{(i:n)}^{*2} + \sum_{i=1}^{\frac{n}{2}} \sigma_{(n-i+1:n)}^{*2} \right),$$

$$Var(\bar{X}_{PPDRSSO}) = \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n+1}{2}} \sigma_{(i:n)}^{*2} + \sum_{i=1}^{\frac{n-1}{2}} \sigma_{(n-i+1:n)}^{*2} \right).$$

The RE for the symmetric distribution is

$$RE(\bar{X}_{SRS}, \bar{X}_{PPDRSS}) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{PPDRSS})},$$

and $RE(\bar{X}_{SRS}, \bar{X}_{PPDRSS}) = \frac{Var(\bar{X}_{SRS})}{MSE(\bar{X}_{PPDRSS})}$ for asymmetric distribution.

Lemma (i) $Var(\bar{X}_{MRSSO}) \leq Var(\bar{X}_{ERSO})$ for odd sample size (n) .

(ii) $Var(\bar{X}_{MRSSO}) \leq Var(\bar{X}_{QRSSO})$ for odd sample size (n) .

(iii) $Var(\bar{X}_{MRSSO}) \leq Var(\bar{X}_{PRSSO})$ for odd sample size (n) .

(iv) In PRSS if the value of $p = 0.25$, then $Var(\bar{X}_{PRSS}) = Var(\bar{X}_{QRSS})$ for each sample size, whether even or odd.

Proof. To prove **(i)**, using (2.1) and (2.2)

$$\begin{aligned} Var(\bar{X}_{ERSO}) &= \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n-1}{2}} \sigma_{(1:n)}^2 + \sum_{i=\frac{n-1}{2}+1}^{n-1} \sigma_{(n:n)}^2 + \sigma_{(\frac{n+1}{2}:n)}^2 \right) \\ &= \frac{1}{n^2} \left(\left(\frac{n-1}{2} \right) \sigma_{(1)}^2 + \left(\frac{n-1}{2} \right) \sigma_{(n)}^2 + \sigma_{(\frac{n+1}{2}:n)}^2 \right) \\ &= \frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma_{(1)}^2 + \sigma_{(n)}^2) + \frac{1}{n^2} \sigma_{(\frac{n+1}{2}:n)}^2 \\ &= \frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma_{(1)}^2 + \sigma_{(n)}^2) + Var(\bar{X}_{MRSSO}) \end{aligned}$$

Remark that $\frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma_{(1)}^2 + \sigma_{(n)}^2) \geq 0$, then

$$Var(\bar{X}_{MRSSO}) \leq Var(\bar{X}_{ERSO}).$$

(ii) using (2.4) and (2.2),

$$\begin{aligned} Var(\bar{X}_{QRSSO}) &= \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n-1}{2}} \sigma_{(q_1(n+1):n)}^2 + \sum_{i=\frac{n-1}{2}+1}^{n-1} \sigma_{(q_3(n+1):n)}^2 + \sigma_{(\frac{n+1}{2}:n)}^2 \right) \\ &= \frac{1}{n^2} \left(\left(\frac{n-1}{2} \right) \sigma_{(q_1(n+1))}^2 + \left(\frac{n-1}{2} \right) \sigma_{(q_3(n+1))}^2 + \sigma_{(\frac{n+1}{2}:n)}^2 \right) \\ &= \frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma_{(q_1(n+1))}^2 + \sigma_{(q_3(n+1))}^2) + \frac{1}{n^2} \sigma_{(\frac{n+1}{2}:n)}^2 \\ &= \frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma_{(q_1(n+1))}^2 + \sigma_{(q_3(n+1))}^2) + Var(\bar{X}_{MRSSO}) \end{aligned} \quad (2.7)$$

Remark that $\frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma_{(q_1(n+1))}^2 + \sigma_{(q_3(n+1))}^2) \geq 0$, then

$$Var(\bar{X}_{MRSSO}) \leq Var(\bar{X}_{QRSSO}).$$

(iii) using (2.6) and (2.2),

$$\begin{aligned}
 Var(\bar{X}_{PRSSO}) &= \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n-1}{2}} \sigma^2_{(p(n+1):n)} + \sum_{i=\frac{n-1}{2}+1}^{n-1} \sigma^2_{(q(n+1):n)} + \sigma^2_{(\frac{n+1}{2}:n)} \right) \\
 &= \frac{1}{n^2} \left(\left(\frac{n-1}{2} \right) \sigma^2_{(p(n+1))} + \left(\frac{n-1}{2} \right) \sigma^2_{(q(n+1))} + \sigma^2_{(\frac{n+1}{2})} \right) \\
 &= \frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma^2_{(p(n+1))} + \sigma^2_{(q(n+1))}) + \frac{1}{n^2} \sigma^2_{(\frac{n+1}{2})} \\
 &= \frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma^2_{(p(n+1))} + \sigma^2_{(q(n+1))}) + Var(\bar{X}_{MRSSO}) \quad (2.8)
 \end{aligned}$$

Remark that $\frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma^2_{(q_1(n+1))} + \sigma^2_{(q_3(n+1))}) \geq 0$, then

$$Var(\bar{X}_{MRSSO}) \leq Var(\bar{X}_{PRSSO}).$$

(iv) If $p = q_1 = 0.25$ and $q = q_3 = 0.75$ and the sample size (n) is odd,

from (2.7) and (2.8),

$$\begin{aligned}
 Var(\bar{X}_{QRSSO}) - \frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma^2_{(q_1(n+1))} + \sigma^2_{(q_3(n+1))}) &= \\
 Var(\bar{X}_{PRSSO}) - \frac{1}{n^2} \left(\frac{n-1}{2} \right) (\sigma^2_{(p(n+1))} + \sigma^2_{(q(n+1))}) &
 \end{aligned}$$

then

$$Var(\bar{X}_{QRSSE}) = Var(\bar{X}_{PRSSE}),$$

If $p = q_1 = 0.25$ and $q = q_3 = 0.75$ and the sample size (n) is even,

from (2.3) and (2.5),

$$\begin{aligned}
 Var(\bar{X}_{QRSSE}) - \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma^2_{(q_1(n+1):n)} + \sum_{i=\frac{n}{2}+1}^n \sigma^2_{(q_3(n+1):n)} \right) &= \\
 Var(\bar{X}_{PRSSE}) - \frac{1}{n^2} \left(\sum_{i=1}^{\frac{n}{2}} \sigma^2_{(p(n+1):n)} + \sum_{i=\frac{n}{2}+1}^n \sigma^2_{(q(n+1):n)} \right) &
 \end{aligned}$$

then

$$Var(\bar{X}_{QRSSO}) = Var(\bar{X}_{PRSSO}).$$

3. Simulation Study

Based on 30000 replications, a computer simulation is conducted to study the behavior of the efficiency of the sample mean using RSS, PPRSS, ERSS, MRSS, QRSS, PRSS with $p = 0.33$ and 0.8 , DRSS, QDRSS, PPDRSS with respect to SRS. The sample size n should be kept very small so that visual ranking errors will not destroy the gain in efficiency. Random observations are generated from (1) Standard normal distribution with $(\mu = 0, \sigma^2 = 1)$ (2) Beta distribution with $(\alpha = 3, \beta = 3)$ (3) Logistic distribution with $(\alpha = -1, \beta = 1)$ (4) Exponential distribution with $(\lambda = 1)$ (5) Gamma distribution with $(\alpha = 1, \beta = \frac{1}{2})$ (6) Weibull distribution with $(\lambda = 1, k = 2)$, were generated by the (MATHEMATICA) package. Samawi and Muttlak^[20] used 5000 replications in their paper because of that the replications were extended to be more than 5000.

The simulation results are summarized in list of figures from Figure 1 to Figure 6 for the performance of the samples means for the sample sizes $n = 6, 9, 11$ for one cycle ($m = 1$). The results of these figures show the relative efficiency (the ratio of the variances) of the estimators of the mean relative to SRS.

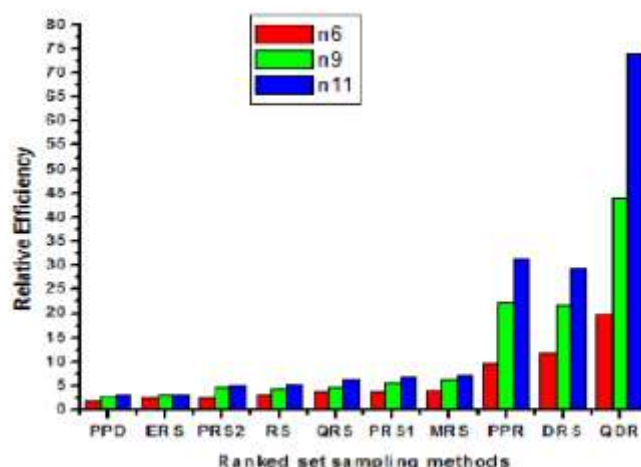


Figure 1 RE for standard normal distribution with ($\mu = 0, \sigma^2 = 1$)

For Standard Normal Distribution:

- Regarding the sample size n , the efficiency of the mean estimators using the ranked set sampling methods increases as n increases for almost results.
- Figure 1 shows that estimating the population mean using QDRSS is substantially more efficient than the other methods and ERSS is less efficient than the other methods.
- For almost sample sizes, MRSS is more efficient than RSS, ERSS, QRSS and PRSS with ($p = 0.33$ and 0.8). While DRSS is more efficient than RSS because it's a two stage of the RSS and give more accurate result.
- The results for RSS, ERSS, MRSS, QRSS, PRSS with ($p = 0.33$ and 0.8) and PPDRSS are very close more than the results from PPRSS, DRSS and QDRSS.
- The population mean estimator using DRSS is more efficient than the population mean estimator using PPDRSS.

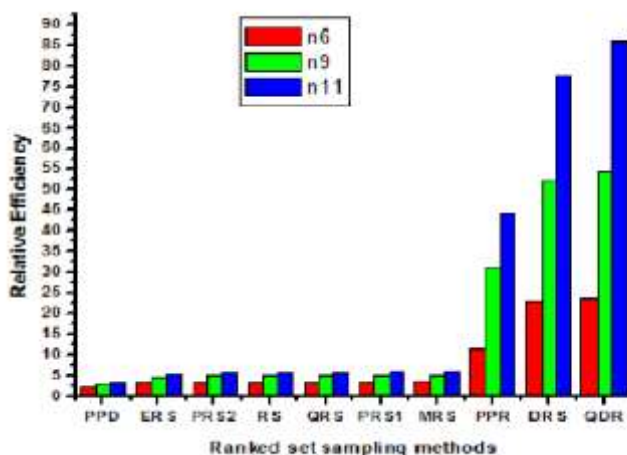


Figure 2 RE for beta distributions with $(\alpha = 3, \beta = 3)$

For Beta Distribution:

- Regarding the sample size n , the efficiency of the mean estimators using the ranked methods increases as n increases for almost results.
- Figure 2 shows that estimating the population mean using QDRSS is more efficient than the other methods and ERS is less efficient than the other methods.
- For all the sample sizes, the results for RSS, ERSS, MRSS, QRSS and PRSS for $(p = 0.33 \text{ and } 0.8)$ are look constantly.
- The population mean estimator using DRSS and QDRSS are more efficient than the population mean estimator using RSS and QRSS respectively.

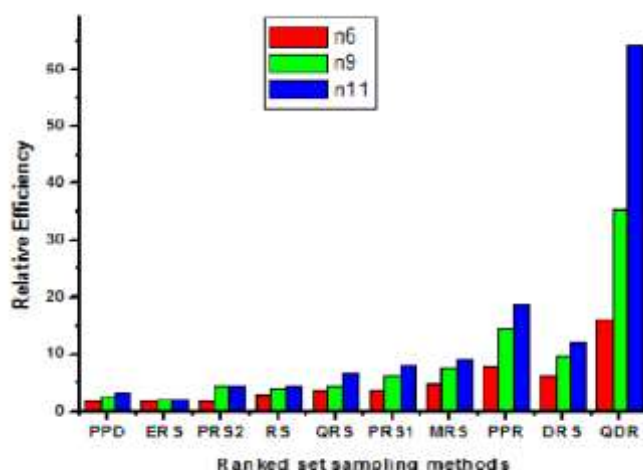
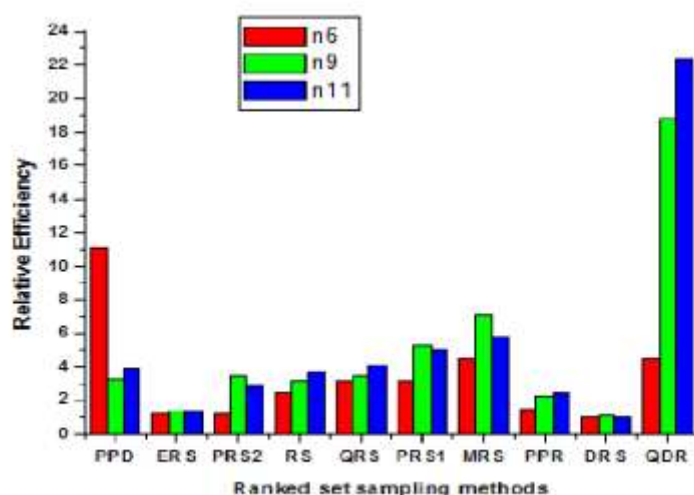


Figure 3 RE for logistic distribution with $(\alpha = -1, \beta = 1)$

For Logistic Distribution:

- Regarding the sample size n , the efficiency of the mean estimators using the ranked methods increases as n increases for almost results.
- Figure 3 shows that estimating the population mean using QDRSS is substantially more efficient than the other methods and ERSS is less efficient than the other methods.
- For almost sample sizes, MRSS is more efficient than RSS, ERSS, QRSS, PRSS with ($p = 0.33$ and 0.8) and PPDRSS. While DRSS is more efficient than RSS because it is a two stage of the RSS and give more accurate result.
- The results for RSS, ERSS, MRSS, QRSS, PRSS with ($p = 0.33$ and 0.8) and PPDRSS are very close more than the results from PPRSS, DRSS and QDRSS.

Figure 4 RE for exponential distributions with ($\lambda = 1$)**For Exponential Distribution:**

- Regarding the sample size n , the efficiency of the mean estimators using the ranked methods increases as n increases for almost results.
- Figure 4 shows that estimating the population mean using QDRSS is substantially more efficient than the other methods and ERSS is less efficient than the other methods.
- For almost sample sizes, MRSS is more efficient than RSS, ERSS, QRSS, PRSS with ($p = 0.33$ and 0.8), PPRSS and DRSS.
- The results for RSS, ERSS, MRSS, QRSS, PRSS for ($p = 0.33$ and 0.8) and PPDRSS are very close more than the results from the other methods.

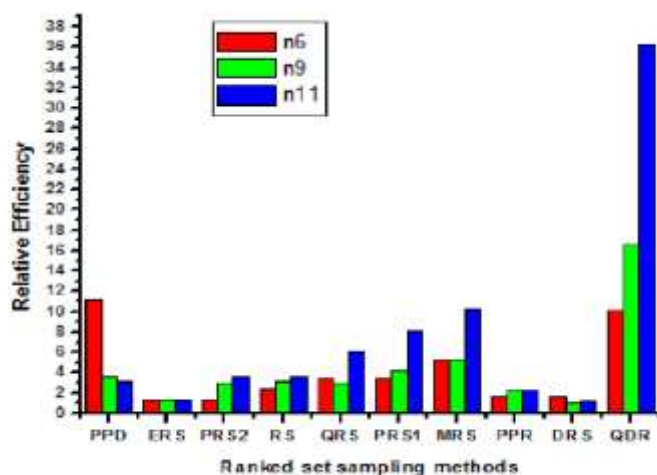


Figure 5 RE for gamma distribution with $(\alpha = 1, \beta = \frac{1}{2})$

For Gamma Distribution:

- This figure looks like figure 4 for all the methods because gamma distribution with shape parameter $\alpha = 1$ and scale parameter β is an exponential (β) distribution.
- Figure 5 shows that estimating the population mean using QDRSS is substantially more efficient than the other methods and ERSS is less efficient than the other methods.
- For almost sample sizes, MRSS is more efficient than RSS, ERSS, QRSS, PRSS with ($p = 0.33$ and 0.8), PPRSS and DRSS.
- The results for RSS, ERSS, MRSS, QRSS, PRSS for ($p = 0.33$ and 0.8) and PPDRSS are very close more than the results from the other methods.

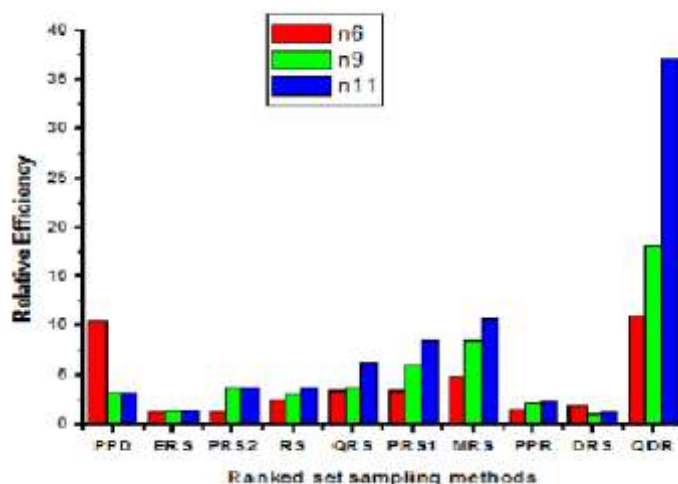


Figure 6 RE for weibull distribution with $(\lambda = 1, k = 2)$

For Weibull Distribution:

- This figure looks like figure 4 and figure 5 for all the methods because weibull $(1, k)$ random variable is an exponential random variable with mean k . Both the gamma and weibull distributions can be seen as generalizations of the exponential distribution.
- Figure 5 shows that estimating the population mean using QDRSS is substantially more efficient than the other methods and ERSS is less efficient than the other methods.
- For almost sample sizes, MRSS is more efficient than RSS, ERSS, QRSS, PRSS with $(p = 0.33 \text{ and } 0.8)$, PPRSS and DRSS.
- The results for RSS, ERSS, MRSS, QRSS, PRSS for $(p = 0.33 \text{ and } 0.8)$ and PPDRSS are very close more than the results from the other methods.

4. Summary and Conclusions

In this article, A comparative study for RE of the sample mean using RSS, PPRSS, ERSS, MRSS, QRSS, PRSS with $(p = 0.33 \text{ and } 0.8)$, DRSS, QDRSS and PPDRSS with respect to SRS is discussed. Based on the simulation results from the comparative study, the findings may be summarized as follows:

1. Regarding the sample size n , the efficiency of the mean estimators using the ranked set sampling methods increasing as the sample size n increasing, especially appear in symmetric distributions.
2. All the ranked set sampling methods are more efficient than SRS, the results given in list of figures confirmed this fact.
3. The QDRSS method is more efficient than the other methods which can called the best method for estimating the population mean.
4. The ERSS method is less efficient than the other methods which can called the worst method for estimating the population mean, especially appear in asymmetric distributions.

References

1. Al-Saleh, M. F., and Al- Kadiri, M. A., (2000). *Double ranked set sampling*. Statistics and Probability Letters 48, 205-212.

2. Al-Saleh , M. F., and Al- Omari, A. I., (2009). *Quartile double ranked set sampling for estimating the population mean*. Economic Quality control. 24, 243-253.
3. Balakrishnan, N., and Chen, W .W., (1997). *CRC Handbook of Tables for Order Statistics from Inverse Gussian Distributions with Applications*, CRC Press, Boca Raton.
4. Biradar, B. S., and Santosha, C. D., (2015). *Estimation of the Population Mean Based on Extremes Ranked Set Sampling*. American Journal of Mathematics and Statistics 5, 32-36.
5. Dell, T. R. and Clutter , J. L., (1972). *Ranked set sampling theory with order statistics background*. Biometrics 28, 545-555.
6. Haq, A., Brown, J., Moltchanova, E., and Al-Omari , A. I., (2013). *Partial ranked set sampling design*. Environmetrics 24, 201-207.
7. Haq, A., Brown, J., Moltchanova, E., and Al-Omari , A. I., (2016). *Paired double-ranked set sampling*. Communications in Statistics - Theory and Methods 10, 2873-2889.
8. Harter, H. L. and Balakrishnan, N., (1996). *CRC Handbook of Tables for the use of Order Statistics in Estimation*. CRC Press, Boca Raton.
9. Ibrahim, A., Al-Omari, A. I. , and Khalifa H. J., (2008). *Percentile Double Ranked Set Sampling*. Journal of Mathematics and Statistics 4, 60-64.
10. McIntyre, G. A., (1952). *A method of unbiased selective sampling, using ranked sets*. Australian J. Agricultural Research 3, 385-390.
11. Muttalak, H.A., (1996). *Pair ranked set sampling*. The Biometrical Journal 38, 879-885.
12. Muttalak, H. A., (1997). *Median ranked set sampling*. Journal of Applied Statistical Sciences 6, 245-255.
13. Muttalak, H. A., (1998). *Median ranked set sampling with concomitant variables and a comparison with ranked set sampling and regression estimator*. Environmetrics 9,255-267.

14. Muttalak, H. A., (2003). *Investigating the use of quartile ranked set samples for estimating the population mean*. Journal of Applied Mathematics and Computation 146,437-443.
15. Muttalak, H. A., (2003). *Modified ranked set sampling methods*. Pakistan Journal of Statistics 19, 315-323.
16. Patil, G. P., Sinha, A. K., Taillie, C., (1993). *Relative precision of ranked set sampling: Comparison with the regression estimator*. Environmetrics 4, 399-412.
17. Richard, S. L., Dennis, W.,and William, M., (2008). *Mathematical Statistics with Applications* .7th ed. Thomson Higher Education. USE.
18. Samawi, H. M., (2011). *Varied set size ranked set sampling with applications to mean and ratio estimation*. The International Journal of Simulation Modeling 31 , 6–13.
19. Samawi, H. M., Abu-Dayyeh, W. and Ahmed, S., (1996). *Extreme ranked set sampling*. The Biometrical Journal 30 , 577-586.
20. Samawi, H. M., and Muttalak, H. A., (1996). *Estimation of Ratio Using Rank Set Sampling* . Biometrical Journal 38 ,753-764.
21. Takahasi, K. and Wakimoto, K., (1968). *On unbiased estimates of the population mean based on the sample stratified by means of ordering*. Annals of the Institute of Statistical Mathematics 20 , 1-31.

جامعة القاهرة
معهد الدراسات والبحوث الإحصائية

المؤتمر السنوى الثانى والخمسين للإحصاء
وعلم الحاسب وبحوث العمليات

الإحصاء التطبيقي

٢٥-٢٧ ديسمبر ٢٠١٧

فهرس الإحصاء التطبيقى

٢٢-١	١ أثر مشكلة التجميع على جودة توفيق نموذج الإنحدار الخطى " دراسة قياسية لمحددات القيمة السوقية لسعر السهم " مصطفى عبد المنعم الخواجة ، نبيلة أحمد سعد
٣٧-٢٣	٢ أهمية قواعد وإجراءات السلامة والصحة المهنية على مراحل إدارة الإزمات والكوارث (دراسة تحليلية على إحدى شركات الإتصالات) عصام عطية عبد المنعم
٥١-٣٨	٣ مؤشرات إحصائية للحوادث المرورية بإستخدام تحليل السلاسل الزمنية بمنطقة عسير السعودية للفترة من (١٤١٧-١٤٣٤) محمد محمد أحمد المزاح

أثر مشكلة التجميع على جودة توفيق نموذج الانحدار الخطى

" دراسة قياسية لمحددات القيمة السوقية لسعر السهم "

¹مصطفى عبد المنعم محمد الخواجة

²نبيلة احمد سعد على

المستخلص

تعتبر مشكلة التجميع أحد مشاكل القياس التى يتأثر بها نموذج الانحدار الخطى، والتى تنشأ نتيجة الإهتمام بالسلوك التجميعى والحصول على بيانات تجميعية وإستخدامها فى تقدير النموذج وإتخاذ القرارات من خلاله. ويمثل الوصول لأثر مشكلة التجميع على جودة توفيق نموذج الانحدار الخطى أحد أهداف الدراسة الحالية، وفى حين تم استقراء العديد من الدراسات التى لخصت تلك الآثار. وقد خلصت الدراسة إلى بعض تلك الآثار بالتطبيق على محدثات القيمة السوقية لسعر السهم فى سوق الأوراق المالية المصرية.

ومن تلك الآثار انخفاض كفاءة مقدرات OLS، وزيادة قيم معامل التحديد إذا تم تجميع البيانات على مستوى الزمن. ولكن لم يتحدد أثر واضح للتجميع على قيم معامل التحديد إذا تم التجميع على مستوى الوحدات القطاعية. كما يُحسن التجميع على مستوى الزمن من قدرة نموذج الانحدار الخطى على التنبؤ، بينما يضعف التجميع على مستوى الوحدات القطاعية من قدرة النموذج على التنبؤ. وتنخفض دقة تنبؤات النموذج التجميعى الذى يعتمد على المجموع الكلى فى التجميع بينما تزداد دقة التنبؤات التى تعتمد على أحد مقياس النزعة المركزية فى التجميع.

الكلمات الدالية للبحث:

تجميع البيانات، جودة توفيق نموذج الانحدار الخطى، دقة تنبؤات نموذج الانحدار الخطى، قدرة نموذج الانحدار الخطى على التنبؤ، كفاءة مقدرات نموذج الانحدار الخطى.

1- مقدمة

يُطلق مصطلح التجميع Aggregation إذا تم أخذ قيمة واحدة فى السنة كمتوسط، أو إجمالى عدة قيم لمتغير ما عند نقطة زمنية محددة. كما تهتم نظرية التجميع بتحويل العلاقات التفصيلية إلى علاقات تجميعية ودراسة العلاقة بينهما. ويهتم الكثير من الباحثين بوصف العديد من العلاقات الإقتصادية فى

¹د. مصطفى عبد المنعم محمد الخواجة

أستاذ الإحصاء التطبيقى بقسم الإحصاء والرياضة والتأمين
كلية التجارة جامعة الاسكندرية

²نبيلة أحمد سعد على

المعيدة بقسم الإحصاء والرياضة والتأمين
كلية التجارة جامعة الاسكندرية

شكل نموذج تجميعى، والذي يعتمد على بيانات يتم تجميعها على عدة مستويات تختلف وفقاً لنوع التجميع.

وذلك بالرغم من الآثار المختلفة لاستخدام هذا النموذج من حيث تأثيره على كفاءة التقديرات، والقوة التفسيرية للنموذج، والقدرة التنبؤية للنموذج ودقة التنبؤات. وترجع تلك الآثار لأن التعامل مع بيانات تفصيلية يسمح بوجود دقة أكبر فى التوصيف والتقدير والتنبؤ. وتختلف تلك الآثار باختلاف نوع وأسلوب التجميع المتبع. وعليه يمكن تلخيص مشكلة الدراسة فى التساؤلات الآتية:

- 1- ما هو أثر مشكلة التجميع على كفاءة مقدرات نموذج الانحدار الخطى؟.
- 2- بما تؤثر مشكلة التجميع على القوة التفسيرية لنموذج الانحدار الخطى؟.
- 3- كيف تتأثر القدرة التنبؤية لنموذج الانحدار الخطى بمشكلة التجميع، وكذلك دقة التنبؤات؟.

وتهدف الدراسة بصورة أساسية إلى الإجابة عن التساؤلات السابقة من خلال التعرف على مشكلة التجميع، وعرض أنواع التجميع المختلفة كالتجميع على مستوى الزمن والتجميع على مستوى الوحدات القطاعية، وذلك باستخدام المجموع الكلى أو أحد مقاييس النزعة المركزية كأساليب للتجميع. ودراسة آثار التجميع على كفاءة مقدرات نموذج الانحدار الخطى المقدر، والقوة التفسيرية والقدرة التنبؤية للنموذج، وأيضاً دقة تنبؤات النموذج. ويتم ذلك عملياً من خلال دراسة قياسية لبعض محددات القيمة السوقية لسعر السهم فى سوق الأوراق المالية المصرية.

كما تتبع أهمية الدراسة فى أنها :

- 1- تستعرض أهم الدراسات السابقة التى قدمت مشكلة التجميع كأحد مشاكل القياس، وأنواع وأساليب تجميع البيانات داخل العلاقات الخطية.
- 2- تقدم دراسة قياسية لبعض محددات القيمة السوقية لسعر السهم فى سوق الأوراق المالية المصرية، وذلك لدراسة أثر مشكلة التجميع على جودة توفيق نموذج الانحدار الخطى.

2- أنواع التجميع

يمكن تقسيم تجميع البيانات فى العلاقات إلى نوعين هما: تجميع البيانات الخطى، وتجميع البيانات غير الخطى. وفى كلا النوعين يمكن تجميع البيانات على مستوى الزمن Temporal Aggregation، أو تجميع البيانات على مستوى الوحدة القطاعية Longitudinal Aggregation. ويقصد بالوحدة القطاعية "المؤسسة أو القطاع أو البلد أو إقليم ما" ويتوقف ذلك على الظاهرة محل الدراسة.

مع مراعاة أن اختلاف طبيعة البيانات الخاصة بمتغيرات النموذج هي التي تفرض نوع التجميع الذى سيتطلب تطبيقه. أى أن التجميع على مستوى الزمن يمكن أن يُطبق على بيانات السلاسل الزمنية. فى حين أن بيانات القطاعات المستعرضة وبيانات السلاسل القطاعية يمكن أن يُطبق عليها التجميع على مستوى الوحدة القطاعية، والتجميع على مستوى الزمن.

(1-2) التجميع غير الخطى للبيانات

تناول [8] Plassmann التجميع غير الخطى للبيانات حيث يتحقق ذلك النوع من التجميع فى حالتين إحداهما: أن تكون معادلة النموذج التفصيلى غير خطية، والأخرى إذا تم استخدام أساليب التجميع بصورة غير خطية أى تكوين متغيرات النموذج التجميعية كدالة غير خطية. وقد أشار Garderen [1] et al. إلى مشكلة التجميع غير الخطى، حيث أكد أن التجميع غير الخطى يُعد مشكلة معقدة لا تصل إلى حلول مرضية إلا فى ظل شروط قوية تتعلق بشكل وتكوين النموذج غير الخطى. والتي تعتمد على مدى التجانس فى سلوك المؤسسات أو القطاعات محل الظاهرة على المستوى التفصيلى.

كما وضح [8] Plassmann أن تجميع البيانات فى العلاقات غير الخطية يهتم بتحويل النموذج التفصيلى إلى نموذج تجميعى غير خطى، ثم أخذ اللوغاريتم لتلك المتغيرات التجميعية ذلك لتحويله إلى نموذج انحدار تجميعى خطى.

(2-2) التجميع الخطى للبيانات

وفيه بعض المفاهيم الأساسية لنظرية التجميع والتي طرحها كلٌّ من [10] Theil، و Plassmann [8]، و [6] Monteforte، و [5] Lutero. حيث أشار [10] Theil إلى نظرية تجميع البيانات فى العلاقات الخطية و التى تهتم بتحويل العلاقات التفصيلية إلى علاقات تجميعية، ودراسة العلاقات بين النماذج على المستويين التفصيلى والتجميعى.

● **المستوى التفصيلى** ويتكون من عدد m من الوحدات القطاعية محل الظاهرة مثل: البلاد أو الأقاليم أو المناطق الجغرافية أو الأسر أو الشركات وغيرها، وكل مؤسسة يتاح لها عدد T من البيانات. بحيث يتم وصف كل مؤسسة بمتغير تابع واحد وعدة متغيرات تفسيرية، كما يمكن وصف النموذج التفصيلى كما فى المعادلة الآتية:

$$Y_{(mT,1)} = X_{(mT,k)}\beta_{(k,1)} + u_{(mT,1)} \quad (1)$$

حيث إن:

$Y_{(mT,1)}$: يعبر عن متجه عمود للمتغير التابع التفصيلى.

$X_{(mT,k)}$: تعبر عن مصفوفة المتغيرات التفسيرية التفصيلية.

$\beta_{(k,1)}$: يعبر عن متجه المعاملات نتيجة انحدار المتغير التابع التفصيلى على المتغيرات التفسيرية التفصيلية.

$u_{(mT,1)}$: يعبر عن متجه عمود لحد الخطأ العشوائى فى النموذج التفصيلى.

● **المستوى التجميعى** ويتم وصفه بمتغير تابع تجميعى واحد، وعدة متغيرات تفسيرية تجميعية ومن خلالها يتم تكوين مصفوفة المتغيرات التفسيرية التجميعية. كما يمكن وصف النموذج التجميعى بالمعادلة الآتية:

$$Y_{(T,1)} = X_{(T,k)}\beta_{(k,1)} + u_{(T,1)} \quad (2)$$

حيث إن:

$Y_{(T,1)}$: يعبر عن متجه عمود للمتغير التابع التجميعى.

$X_{(T,k)}$: تعبر عن مصفوفة المتغيرات التفسيرية التجميعية.
 $\beta_{(k,1)}$: تعبر عن متجه المعاملات نتيجة انحدار المتغير التابع التجميعي على المتغيرات التفسيرية التجميعية.
 $u_{(T,1)}$: يعبر عن متجه عمود لحد الخطأ العشوائي في النموذج التجميعي.

ويلاحظ من المعادلتين (1)، (2) أن الفارق بين المستويين التفصيلي والتجميعي هو تقليل عدد المشاهدات على المستوى التجميعي وذلك نتيجة أخذ المجموع الكلي أو أحد مقاييس النزعة المركزية. كما أن عدد معالم نموذج الانحدار لا تتأثر بتجميع البيانات، ولكن بالتأكد تختلف قيمة مقدرات تلك المعالم. ويمكن تجميع البيانات على عدة مستويات مختلفة مثل التجميع على مستوى الزمن والتجميع على مستوى الوحدة القطاعية محل الظاهرة، وذلك إذا كان تجميع البيانات خطي أو غير خطي ويمكن تناولهما كالآتي:

(2-3) التجميع على مستوى الزمن

ذكر [5] Lutero أن تطبيق التجميع على مستوى الزمن على المدى الطويل يُعد بمثابة تحويل لبيانات قصيرة المدى تُشاهد كل فترة قصيرة أي (بيانات تتكرر بصورة كبيرة) مثل البيانات الربع سنوية إلى بيانات طويلة المدى تُشاهد كل فترة طويلة أي (بيانات تتكرر بصورة أقل) مثل البيانات السنوية. ويُعرف بالتجميع على مستوى الزمن، واقتصر هذا النوع من التجميع على نماذج وحيدة المتغير، ثم توسعت الدراسات السابقة إلى نماذج متعددة المتغيرات.

وقد أشار [3] Hassler إلى وجود شكلين من بيانات السلاسل الزمنية التي يُجرى عليها التجميع على مستوى الزمن وهما: بيانات الأوراق المالية Stock Data، وبيانات التدفق Flow Data. وتُعتبر بيانات التدفق عن الاستهلاك الشهري حيث يتم تطبيق التجميع على مستوى الزمن بتحويل البيانات الشهرية إلى بيانات ربع سنوية أو بيانات سنوية . كما يُقصد ببيانات الأوراق المالية إنها الأسعار اليومية للأوراق المالية ومعدلات الصرف. وبتطبيق التجميع على مستوى الزمن يتم تحويل البيانات اليومية إلى بيانات أسبوعية أو بيانات شهرية، وذلك بأخذ أحد مقاييس النزعة المركزية لكل الأيام في الأسبوع للحصول على بيانات أسبوعية، أو أخذ مقاييس النزعة المركزية لكل الأيام في الشهر للحصول على بيانات شهرية.

وقد أكد [7] Petkovic & Veredas أن التجميع على مستوى الزمن لا يؤثر على توصيف النموذج في حالة النماذج الساكنة. إلا أنه يؤثر على تكوين وهيكل النموذج في حالة النماذج الديناميكية. وأكد [7] Petkovic & Veredas أن التجميع على مستوى الزمن بصفة عامة ينتج عنه مشكلة ارتباط ذاتي بين البواقي. كما استنتج أيضاً أن التجميع على مستوى الزمن يؤثر على خصائص المقدرات في النماذج الخطية. وذلك سواء كانت النماذج ساكنة أو ديناميكية. كما يُسبب التجميع على مستوى الزمن فقد في المعلومات والمشاهدات مما يقلل من كفاءة مقدرات النموذج التجميعي. وهذا يؤثر على مدى معقولية وقبول النموذج التجميعي.

وقد درس [9] Rossana & Seater أثر التجميع على مستوى الزمن على خصائص وثبات السلاسل الزمنية، وقد استنتج أن هذا النوع من التجميع يغير من خصائص السلاسل الزمنية للبيانات. ولذلك فإن الاختلافات الأساسية في البيانات التفصيلية تفقد كلياً مما يغير من خصائص السلسلة

الزمنية. أما عن الثبات فقد ذكر أن البيانات التجميعية تعد أكثر ثباتاً على المدى الطويل من البيانات التفصيلية.

ولكن حين تناول Hassler [3] الثبات في ظل مشكلة التجميع على مستوى الزمن تعارضت نتائج مع نتائج Rossana & Seater [9]. حيث أكد أن ثبات السلسلة الزمنية يختلف باختلاف مستوى التجميع، وقد أشار إلى عدم وجود أمر يقيني إن كان الثبات يتزايد أم يتناقص بسبب التجميع على مستوى الزمن. ولكنه أكد على أن ثبات السلسلة الزمنية يقل بزيادة الفترة التي يتم التجميع على أساسها.

وقد أشارت دراسة Gordevicius et al. [2] أن التجميع على مستوى الزمن يُعد معالجة هامة لتلخيص بيانات السلاسل الزمنية. وتتوافر بيانات سلاسل زمنية في معظم القطاعات كالقطاع المالي والقطاع الطبي وغيرها من القطاعات. كما يُستخدم التجميع على مستوى الزمن لتلخيص مجموعات كبيرة من البيانات بقيمة محددة عند نقطة زمنية معينة أو عند فترة زمنية معينة.

(2-4) التجميع على مستوى الوحدة القطاعية

لقد عرف Lutero [5] التجميع على مستوى الوحدة القطاعية أنه ذلك النوع الذي يهتم بالتجميع عبر النطاق الجغرافي، وعادة يكون على مستوى أقاليم أو مناطق إدارية أو مجموعة من المؤسسات أو القطاعات مثل الأسر أو الشركات وذلك باستخدام أحد أساليب التجميع. ويتم تناول و دراسة أساليب التجميع في المبحث التالي لما له من أهمية قصوى حيث لا يمكن الاقتصار على أسلوب بعينه للتعامل مع كافة أنواع البيانات، كما ينعكس أثر التجميع على خصائص نموذج الانحدار الخطي المقدر بتغير أسلوب التجميع.

وبدأ [10] Theil التحدث عن مشكلة التجميع على مستوى المؤسسات في مجال الاقتصاد القياسي كأحد مشاكل القياس، وقد بدأ بافتراضين وهما: أن النموذجين التفصيلي والتجميعي يتم توصيفهما توصيفاً صحيحاً أمثلاً، وأن كل معادلة تفصيلية في النموذج التفصيلي لها نفس العدد من المتغيرات التفسيرية. وقد ذكر [7] Petkovic & Veredas أن هذا النوع من التجميع يتم لنماذج السلاسل القطاعية التي يتم مشاهدتها على أساس بعدين أحدهم الزمن والآخر الوحدات القطاعية. لذا يترتب على هذا النوع من التجميع ظهور مشكلة عدم ثبات التباين.

3- أساليب التجميع

بالعودة إلى مصطلح التجميع يمكن استنتاج أن أساليب التجميع متعددة كالتجميع باستخدام المجموع الكلي وذلك كما أشار [6] Monteforte، أو الاعتماد على أحد مقاييس النزعة المركزية كالوسط الحسابي كما عرض [10] Theil، والوسط التوافقي وذلك كما ذكر [1] Garderen et al، واستخدام الوسط المرجح بأوزان لعدد من السلاسل التفصيلية وهذا ما نوه عنه [8] Plassmann. وعليه يمكن أن تنقسم أساليب التجميع إلى المجموع الكلي أو استخدام أحد مقاييس النزعة المركزية.

وأكد [11] Yager أن أساليب التعامل مع البيانات تختلف وفقاً لنوع البيانات أي أنه إذا كانت البيانات كمية يصلح لها المجموع الكلي والوسط الحسابي والوسيط والمنوال كأساليب للتجميع. ولكن إذا كانت البيانات كمية ولها أوزان معينة خاصة بها فإن الأسلوب المناسب هو الوسط المرجح

بأوزان. أما إذا كانت البيانات غير كمية "وصفية" فإنه لا يمكن استخدام أى مقياس للنزعة المركزية سوى المنوال فى تلك الحالة. ويتم عرض بعض أساليب التجميع كما يلي:

(3-1) الوسط الحسابى للمشاهدات كأحد أساليب التجميع

يُعد [10] Theil أول من أشار إلى التجميع الخطى للبيانات، وفى ظل فروض [10] Theil بأن النموذجين التفصيلى والتجميعى يتم توصيفهما توصيفاً صحيحاً أمثلاً، وأن كل معادلة تفصيلية فى النموذج التفصيلى لها نفس العدد من المتغيرات التفسيرية. وبالرجوع إلى المعادلة التفصيلية (1)، وبتطبيق التجميع الخطى على مستوى الوحدات القطاعية وذلك بتكوين المتغير التجميعى باستخدام الوسط الحسابى كأحد مقاييس النزعة المركزية كالاتى:

$$\bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i \quad (3)$$

وبالتعويض من معادلة (1) فى معادلة (3)، فإن النموذج التجميعى يأخذ الشكل الآتى:

$$\bar{Y} = \frac{1}{m} \sum_{i=1}^m X_i \beta_i + \frac{1}{m} \sum_{i=1}^m u_i \quad (4)$$

وفى حالة إتاحة البيانات فى صورتها التجميعية، فإنه يتم التعامل مع متغيرات تجميعية $(\bar{X}, \bar{Y})$ أى التعامل مع نموذج تجميعى به علاقة خطية بسيطة فى متغيرات تجميعية. ومتجه معاملتهما يناظر متجه المعاملات فى النموذج التفصيلى، والبواقي فى العلاقة التجميعية تساوى متوسط البواقي على المستوى التفصيلى المناظرة لها. ويتم التعامل مع المعادلة التجميعية التى تأخذ الشكل الآتى:

$$\bar{Y} = \bar{X} \beta + \frac{1}{m} \sum_{i=1}^m u_i \quad (5)$$

وبناءً على ما سبق فإن المعادلة التجميعية الصحيحة (4) لن تساوى المعادلة المقدرة (5) إلا إذا كانت المعاملات على المستوى التفصيلى متساوية، بينما إذا كانت المعاملات على المستوى التفصيلى غير متساوية سيكون هناك خطأ فى التوصيف. ولذلك يذكر [10] Theil أن التجميع يعد حالة خاصة وتطبيقاً لتحليل التوصيف Specification analysis.

وقد استعان [10] Theil بإجراء الانحدار المساعد Auxiliary Regression، ذلك النوع من الانحدار الذى يستخدم لدراسة التناظر بين متغيرات النموذجين التفصيلى والتجميعى، وفيه يتم إجراء انحدار للمتغير التفسيري التفصيلى مضروباً فى مقلوب عدد المؤسسات $(\frac{1}{m} X_i)$ على المتغير التفسيري التجميعى $\bar{X}$. ويعبر عن معادلة الانحدار المساعد بصيغة المصفوفات بالشكل الآتى:

$$\frac{1}{m} X_i = \bar{X} B_i + \epsilon_i \quad (6)$$

حيث تعبر B_i عن مصفوفة معاملات نموذج الانحدار المساعد، وتأخذ المعادلة الآتية:

$$B_i = (\bar{X} \bar{X})^{-1} \bar{X} \frac{1}{m} X_i \quad (7)$$

وهى مصفوفة الأوزان الخاصة بالمعاملات المناظرة وغير المناظرة بين النموذجين التفصيلى والتجميعى. ويكون فيها مجموع الأوزان الخاصة بالمعاملات المناظرة يساوى الواحد الصحيح، ومجموع الأوزان الخاصة بالمعاملات غير المناظرة يساوى صفر. ويُقصد بالتناظر هنا على سبيل المثال أن توقع مقدر معلمة الميل على المستوى التجميعى يعتمد على معالم الميل على مستوى جميع

المؤسسات أو القطاعات (المستوى التفصيلى) فى حين أن مقدرات معالم النموذج التجميعى الأخرى لا تتأثر بمعالم الميل على مستوى جميع المؤسسات أو القطاعات.

(2-3) المجموع الكلى للمشاهدات كأحد أساليب التجميع

لقد فرض [5] Lutero أن معادلة النموذج التفصيلى بصيغة المصفوفات للمؤسسة $i = 1, \dots, m$ ، وذلك عبر نقاط زمنية $t = 1, \dots, T$ تأخذ المعادلة (1). وبفرض أن التجميع الذى يتم تطبيقه تجميع خطى على مستوى المؤسسات أو القطاعات، وبتكوين المتغير التابع التجميعى باستخدام المجموع الكلى لمشاهدات المتغيرات التابعة على مستوى عدد m من المؤسسات بالشكل الآتى:

$$Y_{at} = \sum_{i=1}^m Y_{it} \quad (9)$$

حيث إن Y_{at} يعبر عن إجمالى قيم مشاهدات المتغير التابع وذلك على مستوى جميع المؤسسات محل الدراسة وعند نقطة زمنية t .

وبتكوين المتغير التفسيري التجميعى باستخدام المجموع الكلى لمشاهدات المتغيرات التفسيرية التفصيلية كما بالشكل الآتى:

$$X_{at} = \sum_{i=1}^m X_{it} \quad (10)$$

حيث X_{at} يعبر عن إجمالى قيم مشاهدات المتغير المستقل وذلك على مستوى جميع المؤسسات محل الدراسة وعند نقطة زمنية t .

فإن النموذج التجميعى المقدر يأخذ المعادلة (2)، بحيث تعرف متغيرات النموذج التجميعى باستخدام المجموع الكلى. ولدراسة العلاقة بين النموذجين التفصيلى والتجميعى والتناظر الإحصائى بينهما يتم الاستعانة بالانحدار المساعد Auxiliary Regression، وذلك بإجراء انحدار لمصفوفة المتغيرات التفسيرية فى النموذج التفصيلى X_{it} لكل مؤسسة على حدة، وذلك الانحدار على المصفوفة القطرية لمتغيرات النموذج التجميعى التفسيرية وكما سبق بالمعادلة (6). ولكن فى ظل النموذج التجميعى الذى تعرف متغيراته بالمجموع الكلى. وتأخذ معادلة الانحدار المساعد الشكل الآتى:

$$X_{it} = X_{at} \alpha_i + \epsilon_{it} \quad (11)$$

بحيث يعبر ϵ_{it} عن حد الخطأ العشوائى لمعادلة الانحدار المساعد، ويعبر α_i عن مصفوفة الأوزان الخاصة بالمعاملات المناظرة وغير المناظرة. ويكون مجموع تلك الأوزان الخاصة بالمعاملات المناظرة يساوى الواحد الصحيح فى حين أن مجموع تلك الأوزان يساوى الصفر للمعاملات غير المناظرة.

وللوصول إلى العلاقة بين حد الخطأ العشوائى على المستوى التفصيلى، وحد الخطأ العشوائى على المستوى التجميعى يتم التعويض من المعادلة (11) فى المعادلة (1) وبأخذ المجموع الكلى على مستوى المؤسسات أو القطاعات، ثم تقدير حد الخطأ العشوائى فى النموذج التجميعى بالشكل التالى:

$$\hat{u}_{at} = \sum_{i=1}^m (\epsilon_{it} \beta_i + \hat{u}_{it}) \quad (12)$$

أى أن حد الخطأ العشوائى فى النموذج التجميعى عبارة عن مجموع مقدارين حد الخطأ العشوائى فى النموذج التفصيلى بعد التجميع، بالإضافة إلى خطأ التجميع وهو حاصل ضرب بواقي نموذج الانحدار المساعد ومقدر معلمة النموذج التفصيلى المناظرة له.

4- دراسة قياسية لبعض محددات القيمة السوقية لسعر السهم

لدراسة الآثار المترتبة على مشكلة التجميع، فقد تتضمن هذا الجزء اطار وبيانات الدراسة القياسية والأساليب الاحصائية المستخدمة ونتائج تلك الدراسة ومناقشة أهم النتائج والدلالات والمضامين الأكاديمية الخاصة بها.

(1-4) متغيرات وبيانات الدراسة

تشتمل الدراسة على ثلاثة متغيرات أحدهم يمثل المتغير التابع وهو القيمة السوقية لسعر السهم اليومى، والمتغيرين الآخرين هما التفسيرين ويمثلان أعلى سعر للسهم تم التعامل اليومى به وأقل سعر أيضاً. وتأخذ معادلة نموذج الانحدار المقدرة الصورة التالية:

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 X_{it} + \hat{\beta}_2 Z_{it}$$

بحيث أن:

$\hat{Y}_{it}$: يمثل القيمة المقدرة لسعر اقبال السهم وهو آخر سعر تم التعامل اليومى به.

X_{it} : يمثل أعلى سعر يومى للسهم تم التعامل به، وهو أحد المتغيرات التفسيرية فى نموذج الانحدار.

Z_{it} : يمثل أقل سعر يومى للسهم تم التعامل به، وهو أحد المتغيرات التفسيرية فى نموذج الانحدار.

وكل من " $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ " يمثلوا مقدرات معالم نموذج الانحدار الخطى.

أما عن بيانات الدراسة تم الاستعانة ببيانات سلاسل زمنية الخاصة بعشر أوراق مالية الأعلى من حيث السيولة والنشاط كما فى جدول (1) ووفقاً للأسبوع الأخير من شهر مارس لعام 2017، حيث يطبق على تلك البيانات التجميع على مستوى الزمن " شهرياً وربع سنوي ". وتجمع تلك البيانات فى صورة سلاسل زمنية عن الفترة الممتدة من أول يناير 1998 إلى آخر يناير 2015 بشكل يومى بعد استبعاد الاجازات الرسمية وأيام العطلة والأيام التى لا يصدر عنها تقرير يومى.

جدول (1)

بيان بالشركات التى يتم بواسطتها تطبيق التجميع على مستوى الزمن

اسم الشركة محل الدراسة	عدد المشاهدات على المستوى التفصيلي	عدد المشاهدات على مستوى الشهر	عدد المشاهدات على المستوى الربع سنوى
1- طلعت مصطفى	1717	86	29
2- مصر الجديدة	3404	205	68
3- السادس من أكتوبر	3771	199	67
4- هيرمس القابضة	3554	187	63
5- نصر للإسكان	4057	204	68
6- أجواء للصناعات	1296	103	35
7-النساجون الشرقيون	3873	203	68
8- جلوبال تيلكوم	3463	172	58
9- المصرية للاتصالات	2196	109	37
10- البنك التجارى الدولى	4138	204	68

ولتطبيق التجميع على مستوى الوحدة القطاعية كالمؤسسات والشركات تم الإستعانة ببيانات فى صورة سلاسل مقطعية لعدد 14 قطاعاً إقتصادياً داخل سوق المقصورة ويحتوى جدول (2) على بيان بتلك البيانات، بحيث يحتوى كل قطاع على عدد من المؤسسات.

جدول (2)

بيان بالشركات التى يتم بواسطتها تطبيق التجميع على مستوى المؤسسات

اسم القطاع محل الدراسة	عدد المؤسسات	عدد المشاهدات على المستوى التفصيلي	عدد المشاهدات على مستوى التجميعي
1-البنوك	12	22222	1940
2- الموارد الأساسية	7	12949	1940
3-الاتصالات	4	7617	1940
4-الأغذية والمشروبات	20	35337	2051
5-التشيد والبناء	19	34025	2051
6-الخدمات المالية	8	15492	2058
7-السياحة والترفيه	9	10339	1940
8-العقارات	22	38677	1940
9-الغاز والبترول	2	3612	1940
10-الكىماويات	6	11296	2036
11-الخدمات والمنتجات الصناعية	11	20179	1940
12-الرعاية الصحية والأدوية	7	10195	2067
13-المنتجات المنزلية والشخصية	7	13139	1940
14-الموزعين وتجار التجزئة	3	3457	1939

(2-4) نتائج الدراسة

تم تجميع وتحليل البيانات باستخدام برمجة R وذلك مستعيناً بالأوامر الهامة كما ذكرها Kabacoff [4]. وتنقسم نتائج الدراسة إلى قسمين رئيسيين يتلخصان فى أثر مشكلة التجميع على مستوى الزمن وأثر مشكلة التجميع على مستوى الوحدات القطاعية، وذلك على كل من: القوة التفسيرية لنموذج الانحدار الخطى، وكفاءة مقدرات معالم النموذج، والقدرة التنبؤية لنموذج الانحدار الخطى، وأيضاً دقة تنبؤات نموذج الانحدار الخطى.

(1-2-4) أثر مشكلة التجميع على مستوى الزمن

أولاً: **القوة التفسيرية للنموذج** وتقاس تلك القوة التفسيرية لنموذج الانحدار الخطى المقدر باستخدام معامل التحديد، تلك القيمة التى تفسر نسبة التغيرات التى تحدث فى المتغير التابع، والتى يمكن إرجاعها إلى المتغيرات المستقلة محل الدراسة. ويمكن الحصول على معامل التحديد فى نموذج الانحدار الخطى من المعادلة الآتية:

$$R^2 = 1 - \frac{\sum_{t=1}^n u_t^2}{\sum_{t=1}^n y_t^2} \quad (13)$$

بحيث إن:

$$\sum_{t=1}^n u_t^2 : \text{يعبر عن مجموع مربعات البواقي فى نموذج الانحدار الخطى.}$$

$$\sum_{t=1}^n y_t^2 : \text{يعبر عن مجموع المربعات الكلية فى نموذج الانحدار الخطى.}$$

فى حين n تعبر عن حجم العينة وتساوى mT فى النموذج التفصيلى أى قبل تجميع البيانات، وتساوى T فى النموذج التجميعى. وبتطبيق التجميع على مستوى الزمن "شهرياً وربع سنوياً" باستخدام بيانات العشر شركات الأكثر نشاطاً وسيولة فى سوق البورصة المصرية، ثم حساب قيمة معامل التحديد فى النموذج التفصيلى، والقيمة المناظرة لها فى النموذج التجميعى وذلك كما بالمعادلة (13). وقد تم الحصول على النتائج المفروغة فى جدول (3) كما يلى.

جدول (3)

معامل تحديد نموذج الانحدار الخطى المتعدد " R^2 " فى ظل التجميع على مستوى الزمن

تجميع البيانات ربع سنوياً			تجميع البيانات شهرياً			نوع التجميع	
المجموع الكلى	الوسيط	الوسط الحسابى	المجموع الكلى	الوسيط	الوسط الحسابى	أسلوب التجميع	الشركات
						النموذج التفصيلى	
100%	99.99%	100%	100%	99.98%	99.99%	99.97%	1- طلعت مصطفى
100%	100%	100%	100%	99.99%	100%	99.99%	2- مصر الجديدة
100%	99.99%	100%	100%	100%	100%	99.98%	3- السادس من أكتوبر
100%	99.99%	100%	100%	99.99%	100%	99.98%	4- هيرمس القابضة
100%	100%	100%	100%	100%	100%	99.94%	5- نصر للإسكان
100%	99.98%	99.99%	100%	99.96%	100%	99.89%	6- أجواء للصناعات
99.94%	99.99%	99.92%	99.85%	99.98%	99.85%	97.98%	7- النساجون الشرقيون
100%	100%	100%	100%	99.99%	100%	99.98%	8- جلوبال تيلكوم
100%	99.97%	99.99%	99.99%	99.90%	99.97%	99.80%	9- المصرية للاتصالات
100%	100%	100%	100%	100%	100%	99.99%	10- البنك التجارى الدولى

وبناءً على الجدول السابق، فإن تجميع البيانات على مستوى الزمن سواءً تجميع على مستوى الشهر أو تجميعاً على المستوى الربع سنوياً، وبواسطة أي من أساليب التجميع كالمجموع الكلى أو أحد مقاييس النزعة المركزية كالوسط الحسابى أو الوسيط، فإن قيمة معامل التحديد فى النموذج التجميعى تزداد بالمقارنة بقيمة معامل التحديد المناظرة لها فى النموذج التفصيلى.

ثانياً: كفاءة مقدرات معالم النموذج ولدراسة أثر مشكلة التجميع على مستوى الزمن (شهرياً وربع سنوياً) على كفاءة مقدرات معالم النموذج يتم استخدام الخطأ المعياري كمقياس لكفاءة المقدرات. ويتم تلخيص ذلك الأثر فى حالة نموذج الانحدار الخطى المتعدد كما هو موضح فى جدول (4) التالى.

جدول (4): الخطأ المعياري لمقدرات معالم نموذج الانحدار الخطى وتأثره بالتجميع على مستوى الزمن

تابع جدول (4)

تجميع البيانات ربع سنوياً			تجميع البيانات شهرياً			نوع التجميع		الشركات
المجموع الكلى	الوسيط	الوسط الحسابى	المجموع الكلى	الوسيط	الوسط الحسابى	أسلوب التجميع	نموذج الانحدار الخطى المقدر	
						النموذج التفصيلى		
0.5910	0.0337	0.0107	0.0984	0.0344	0.0268	0.0088	قيمة المقدر (β_0)	طلعت مصطفى
0.3992	0.0135	0.0063	0.0865	0.0105	0.0056	0.0029	الخطأ المعياري للمقدر (β_0)	
0.5512	0.6354	0.5513	0.5387	0.6404	0.6774	0.5489	قيمة المقدر (β_1)	
0.0288	0.0482	0.0284	0.0174	0.0214	0.0130	0.0061	الخطأ المعياري للمقدر (β_1)	
0.4489	0.3629	0.4487	0.4616	0.3584	0.3247	0.4514	قيمة المقدر (β_2)	
0.0282	0.0473	0.0277	0.0171	0.0207	0.0126	0.0059	الخطأ المعياري للمقدر (β_2)	
0.3802	-0.0002	0.0007	0.1223	-0.0045	0.0081	0.0076	قيمة المقدر (β_0)	مصر الجديدة
0.3944	0.0156	0.0072	0.0959	0.0140	0.0055	0.0059	الخطأ المعياري للمقدر (β_0)	
0.4929	0.5229	0.4956	0.4879	0.6518	0.4929	0.4876	قيمة المقدر (β_1)	
0.0117	0.0259	0.0121	0.0059	0.0187	0.0067	0.0043	الخطأ المعياري للمقدر (β_1)	
0.5062	0.4772	0.5036	0.5110	0.3533	0.5062	0.5113	قيمة المقدر (β_2)	
0.0112	0.0248	0.0116	0.0056	0.0179	0.0063	0.0041	الخطأ المعياري للمقدر (β_2)	
-0.2024	-0.0049	-0.0056	-0.0683	0.0068	-0.0055	-0.0071	قيمة المقدر (β_0)	السادس من أكتوبر
0.3297	0.0134	0.0064	0.0671	0.0057	0.0038	0.0025	الخطأ المعياري للمقدر (β_0)	
0.4917	0.5634	0.4954	0.4891	0.5246	0.4904	0.5101	قيمة المقدر (β_1)	
0.0334	0.0395	0.0359	0.0166	0.0192	0.0170	0.0060	الخطأ المعياري للمقدر (β_1)	
0.5086	0.4397	0.5052	0.5111	0.4749	0.5101	0.4910	قيمة المقدر (β_2)	
0.5086	0.0379	0.0346	0.0160	0.0185	0.0164	0.0058	الخطأ المعياري للمقدر (β_2)	
0.1024	-0.0198	0.0020	0.0033	-0.0044	-0.0009	-0.0041	قيمة المقدر (β_0)	هيرمس القابضة
0.1645	0.0209	0.0038	0.0565	0.0116	0.0036	0.0034	الخطأ المعياري للمقدر (β_0)	
0.4919	0.5187	0.4929	0.4985	0.5159	0.4998	0.5095	قيمة المقدر (β_1)	
0.0057	0.0335	0.0066	0.0054	0.0177	0.0058	0.0039	الخطأ المعياري للمقدر (β_1)	
0.5076	0.4833	0.5067	0.5018	0.4842	0.5003	0.4911	قيمة المقدر (β_2)	
0.0055	0.03184	0.0062	0.0051	0.0168	0.0055	0.0037	الخطأ المعياري للمقدر (β_2)	
0.0736	-0.0048	0.0008	0.0361	0.0112	0.0015	0.0006	قيمة المقدر (β_0)	نصر للاسكان
0.3804	0.0101	0.0063	0.1273	0.0072	0.0063	0.0058	الخطأ المعياري للمقدر (β_0)	
0.5102	0.6126	0.5103	0.4901	0.3761	0.4979	0.5022	قيمة المقدر (β_1)	
0.0176	0.0201	0.0171	0.0144	0.0136	0.0141	0.0085	الخطأ المعياري للمقدر (β_1)	
0.4895	0.3916	0.4895	0.5087	0.6187	0.5012	0.4972	قيمة المقدر (β_2)	
0.0168	0.0191	0.0163	0.0138	0.0130	0.0134	0.0081	الخطأ المعياري للمقدر (β_2)	
0.0227	0.0169	-0.0099	0.0339	-0.0008	-0.0006	0.0075	قيمة المقدر (β_0)	أجواء للصناعات
0.2562	0.0272	0.0078	0.0871	0.0095	0.0071	0.0081	الخطأ المعياري للمقدر (β_0)	
0.6143	0.2662	0.6669	0.5730	0.6899	0.5981	0.5434	قيمة المقدر (β_1)	
0.0197	0.0665	0.0299	0.0144	0.0177	0.0186	0.0101	الخطأ المعياري للمقدر (β_1)	
0.3907	0.7203	0.3437	0.4290	0.3195	0.4067	0.4559	قيمة المقدر (β_2)	
0.0182	0.0605	0.0276	0.0133	0.0164	0.0172	0.0092	الخطأ المعياري للمقدر (β_2)	

الخطأ المعياري لمقدرات معالم نموذج الانحدار الخطى وتأثره بالتجميع على مستوى الزمن

جامعة القاهرة – معهد الدراسات والبحوث الإحصائية

تجميع البيانات ربع سنوياً			تجميع البيانات شهرياً			نوع التجميع		الشركات
المجموع الكلى	الوسيط	الوسط الحسابى	المجموع الكلى	الوسيط	الوسط الحسابى	النموذج التفصيلى	نموذج الانحدار الخطى المقدر	
0.587	0.0119	0.0155	0.3575	0.0114	0.0183	0.0152	قيمة المقدر ($\hat{\beta}_0$)	النساجون الشرقيون
1.155	0.0066	0.0223	0.3469	0.0064	0.0174	0.0146	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
0.3353	0.5242	0.3356	0.3079	0.4647	0.3375	0.3251	قيمة المقدر ($\hat{\beta}_1$)	
0.1152	0.0402	0.1216	0.0886	0.0243	0.0821	0.0373	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
0.6547	0.4749	0.6533	0.6798	0.5329	0.6515	0.6637	قيمة المقدر ($\hat{\beta}_2$)	
0.1106s	0.0389	0.1166	0.0849	0.0232	0.0787	0.0357	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	
0.0479	-0.0001	0.0007	0.0217	0.0021	0.0011	0.0014	قيمة المقدر ($\hat{\beta}_0$)	جلوبال تيلكوم القابضة
0.0449	0.0029	0.0009	0.0199	0.0025	0.0012	0.0009	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
0.5239	0.5095	0.5136	0.5429	0.5253	0.5106	0.5492	قيمة المقدر ($\hat{\beta}_1$)	
0.0121	0.0356	0.0129	0.0125	0.0254	0.0132	0.0056	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
0.4763	0.4912	0.4864	0.4578	0.4743	0.4892	0.4516	قيمة المقدر ($\hat{\beta}_2$)	
0.0118	0.0347	0.0125	0.0122	0.0247	0.0129	0.0055	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	
-0.821	-0.0029	-0.0257	-0.4680	-0.0306	-0.0115	-0.0045	قيمة المقدر ($\hat{\beta}_0$)	المصرية للاتصالات
1.101	0.0469	0.0258	0.3411	0.0475	0.0273	0.0152	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
0.5359	0.5985	0.5494	0.5297	0.5466	0.5415	0.5619	قيمة المقدر ($\hat{\beta}_1$)	
0.0219	0.0455	0.0225	0.0156	0.0252	0.0190	0.0063	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
0.4659	0.4042	0.4537	0.4726	0.4564	0.4604	0.4403	قيمة المقدر ($\hat{\beta}_2$)	
0.0213	0.0446	0.0217	0.0152	0.0242	0.0183	0.0061	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	
-0.2686	0.0051	-0.0018	-0.0994	-0.0065	-0.0001	-0.0061	قيمة المقدر ($\hat{\beta}_0$)	البنك التجارى الدولى
0.2036	0.0111	0.0039	0.0551	0.0058	0.0032	0.0024	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
0.5641	0.6336	0.5980	0.5489	0.5242	0.6251	0.4962	قيمة المقدر ($\hat{\beta}_1$)	
0.0201	0.0425	0.0218	0.0133	0.0164	0.0113	0.0052	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
0.4382	0.3704	0.4049	0.4529	0.4772	0.3785	0.5044	قيمة المقدر ($\hat{\beta}_2$)	
0.0196	0.0415	0.0212	0.0129	0.0160	0.0110	0.0051	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	
0.3944	0.0156	0.0072	0.0959	0.0140	0.0055	0.0059	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
0.4929	0.5229	0.4956	0.4879	0.6518	0.4929	0.4876	قيمة المقدر ($\hat{\beta}_1$)	
0.0117	0.0259	0.0121	0.0059	0.0187	0.0067	0.0043	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
0.5062	0.4772	0.5036	0.5110	0.3533	0.5062	0.5113	قيمة المقدر ($\hat{\beta}_2$)	
0.0112	0.0248	0.0116	0.0056	0.0179	0.0063	0.0041	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	

ويتضح من الجدول السابق زيادة الخطأ المعيارى لمقدرات معالم النموذج التجميعى إذا تم التجميع بأسلوب المجموع الكلى أو أحد مقاييس النزعة المركزية كالوسط الحسابى والوسيط، وذلك بالنسبة لمقدرات معالم الحد الثابت ومقدرات معالم الميل. ولكن هناك إمكانية لانخفاض الخطأ المعيارى لمقدرات معالم النموذج الخطى وذلك فى حالة التجميع باستخدام الوسيط. وذلك كما حدث فى شركة (النساجون الشرقيون للسجاد).

ثالثاً: القدرة التنبؤية لنموذج الانحدار الخطى

يتم تقييم القدرة التنبؤية لنموذج الانحدار الخطى من خلال اختبار مدى استقرار تقديرات معالم النموذج. والذى يعتمد على إضافة بيانات جديدة من خارج العينة محل الدراسة، وتقدير نموذج الانحدار الخطى قبل وبعد إضافة البيانات الجديدة. ثم مقارنة القيم الجديدة المقدرة للمتغير بالقيم

المقدرة التي تم الحصول عليها قبل إضافة البيانات الجديدة. ويتم الاستدلال عن الفرق بين متوسط القيم الفعلية للمتغير التابع بعد الحصول على بيانات جديدة وقبل الحصول على تلك البيانات $(\mu_{Y_2} - \mu_{Y_1})$. ويتم اختبار الفرق بين متوسطي مجتمعين مستقلين حيث يُصاغ فرض العدم كما يلي:

$$H_0: \mu_{Y_2} = \mu_{Y_1}$$

ويتضمن فرض العدم أن هناك استقراراً في تقديرات نموذج الانحدار الخطي. أي أن نموذج الانحدار الخطي له قدرة على التنبؤ حيث لا يوجد فرق بين متوسط القيم الفعلية للمتغير التابع قبل وبعد الحصول على البيانات الجديدة.

ولدراسة تأثير القدرة التنبؤية لنموذج الانحدار الخطي بمشكلة التجميع على مستوى الزمن، تم استخدام العشر شركات محل الاهتمام في الدراسة القياسية. و يوضح جدول (5) أثر التجميع على مستوى الزمن بتجميع البيانات على مستوى الشهر والتجميع الربع سنوياً لتلك البيانات وذلك باستخدام أساليب التجميع المختلفة. وباستخدام القيمة الاحتمالية p-value في اتخاذ القرار بقبول أو رفض فرض العدم محل الاختبار وحيث أن الاختبار ذو طرفين فإن القيمة الاحتمالية هي احتمال أن تكون إحصائية الاختبار المحسوبة أكبر من القيمة المطلقة للقيمة الجدولية المناظرة لها عند مستوى خطأ معين مسموح به وقيمته 0.05.

جدول (5)

نتائج اختبار t لدراسة القدرة التنبؤية لنموذج الانحدار الخطي في ظل التجميع على مستوى الزمن

تجميع البيانات ربع سنوياً			تجميع البيانات شهرياً			نوع التجميع		
المجموع الكلي	الوسيط	الوسط الحسابي	المجموع الكلي	الوسيط	الوسط الحسابي	أسلوب التجميع	معايير الاختبار	الشركات
						النموذج التفصيلي		
26.9	0.326	0.320	7.52	0.380	0.371	0.2494	الفرق بين المتوسطين	طلعت مصطفى
0.68	0.47	0.46	0.90	0.93	0.91	2.79	قيمة "ت" المحسوبة	
0.498*	0.642*	0.648*	0.371*	0.353*	0.366*	0.005	القيمة الاحتمالية	
75	1.59	1.57	28.4	1.33	1.33	1.540	الفرق بين المتوسطين	مصر الجديدة
0.37	0.49	0.49	0.72	0.71	0.71	3.21	قيمة "ت" المحسوبة	
0.711*	0.623*	0.625*	0.474*	0.479*	0.477*	0.001	القيمة الاحتمالية	
2.2	0.11	0.11	2.4	0.107	0.106	0.104	الفرق بين المتوسطين	السادس من أكتوبر
0.03	0.11	0.11	0.18	0.18	0.18	0.74	قيمة "ت" المحسوبة	
0.973*	0.912*	0.915*	0.854*	0.860*	0.861*	0.458*	القيمة الاحتمالية	
5.7	-0.04	-0.04	0.6	-0.003	-0.005	-0.016	الفرق بين المتوسطين	هيرمس القابضة
0.07	-0.03	-0.03	0.04	-0.00	-0.01	-0.09	قيمة "ت" المحسوبة	
0.947*	0.975*	0.975*	0.972*	0.996*	0.995*	0.929*	القيمة الاحتمالية	
25	0.62	0.61	10.5	0.510	0.515	0.521	الفرق بين المتوسطين	نصر للإسكان
0.24	0.38	0.37	0.53	0.53	0.54	2.4	قيمة "ت" المحسوبة	
0.807*	0.704*	0.710*	0.597*	0.595*	0.592*	0.016	القيمة الاحتمالية	
-2.6	-0.14	-0.14	-1.7	-0.139	-0.144	-0.335	الفرق بين المتوسطين	أجواء للصناعات
-0.04	-0.11	-0.11	-0.12	-0.19	-0.19	-1.57	قيمة "ت" المحسوبة	
0.972*	0.915*	0.912*	0.907*	0.853*	0.851*	0.117*	القيمة الاحتمالية	
7.8	0.232	0.216	3.64	0.184	0.175	0.18	الفرق بين المتوسطين	النساجون الشرقيون
0.26	0.49	0.46	0.62	0.66	0.63	2.83	قيمة "ت" المحسوبة	
0.794*	0.627*	0.649*	0.538*	0.508*	0.528*	0.005	القيمة الاحتمالية	
-0.1	-0.037	-0.039	-0.48	-0.024	-0.025	-0.024	الفرق بين المتوسطين	جلوبال تيلكوم القابضة
-0.00	-0.09	-0.09	-0.09	-0.10	-0.10	-0.42	قيمة "ت" المحسوبة	
0.997*	0.932*	0.930*	0.927*	0.924*	0.923*	0.675*	القيمة الاحتمالية	

تابع جدول (5)

نتائج اختبار t لدراسة القدرة التنبؤية لنموذج الانحدار الخطى فى ظل التجميع على مستوى الزمن

الشركات	نوع التجميع	تجميع البيانات شهرياً			تجميع البيانات ربع سنوياً		
		أسلوب التجميع	الوسط الحسابى	الوسيط	الوسط الحسابى	الوسيط	المجموع الكلى
المصرية للاتصالات	الفرق بين المتوسطين	-0.351	-0.357	-0.356	-0.362	-0.365	-12.2
	قيمة "ت" المحسوبة	-4.59	-1.06	-1.05	-0.63	-0.64	-0.29
	القيمة الإحتمالية	0.0001	0.291*	0.294*	0.533*	0.523*	0.771*
البنك التجارى الدولى	الفرق بين المتوسطين	1.442	1.45	1.45	1.81	1.84	80
	قيمة "ت" المحسوبة	5.36	1.19	1.20	0.85	0.86	0.63
	القيمة الإحتمالية	0.0001	0.233*	0.232*	0.399*	0.394*	0.529*

من الجدول السابق يمكن تلخيص بعض النتائج فيما يلى:

- إذا كان النموذج التفصيلى ليس لديه قدرة على التنبؤ، ويوجد عدم استقرار فى تقديرات معالم نموذج الانحدار كما فى بيانات الشركات الآتية: طلعت مصطفى القابضة، ومصر الجديدة للإسكان، ونصر للإسكان والتعمير، والنساجون الشرقيون للسجاد، والمصرية للاتصالات، والبنك التجارى الدولى. فإن القدرة التنبؤية للنموذج التجميعى بجميع أساليبه المجموع الكلى أو أحد مقاييس النزعة المركزية تتحسن.
- إذا كان النموذج التفصيلى لديه القدرة على التنبؤ، ويوجد استقرار فى تقديرات معالم نموذج الانحدار. فإن التجميع على مستوى الزمن يؤدى إلى زيادة القدرة التنبؤية للنموذج التجميعى بجميع أساليبه وذلك كما فى الشركات التالية: شركة جلوبال القابضة، وشركة أجواء للصناعات الغذائية، وشركة هيرمس القابضة، وشركة السادس من أكتوبر.

رابعاً: دقة تنبؤ نموذج الانحدار الخطى

لدراسة دقة تنبؤ نموذج الانحدار الخطى التجميعى تم قياس جذر متوسط مربعات الخطأ. وذلك فى نموذج الانحدار الخطى المتعدد، وعلى المستويين التفصيلى والتجميعى. ذلك النموذج التجميعى الذى يعتمد على الوسط الحسابى والوسيط والمجموع الكلى كأساليب للتجميع. ويتم حساب قيمة جذر متوسط مربعات الخطأ كما بالمعادلة التى يمكن صياغتها بالصورة التالية:

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (y_{it} - \hat{y}_{it})_{it}^2} \quad (15)$$

وبالتجميع على مستوى الزمن "شهرياً وربع سنوياً" لبيانات يومية، وذلك فى نموذج الانحدار المتعدد كما بجدول (6) التالى.

جدول (6)

جذر متوسط مربعات الخطأ RMSE لنموذج الانحدار الخطى المتعدد فى ظل التجميع على مستوى الزمن

نوع التجميع			تجميع البيانات شهرياً			تجميع البيانات ربع سنوياً	
أسلوب التجميع الشركات	النموذج التفصيلي	الوسط الحسابي	الوسط الحسابي	الوسط الحسابي	المجموع الكلى	الوسط الحسابي	المجموع الكلى
طلعت مصطفى القابضة	0.0462	0.0198	0.0373	0.2857	0.0114	0.0249	0.6664
النساجون الشريون للسجاد	0.3928	0.1063	0.0393	2.2698	0.0767	0.0235	4.2619
جلوبال تيلكوم القابضة	0.0309	0.0083	0.0175	0.1425	0.0035	0.0120	0.1859
السادس من أكتوبر	0.0838	0.0274	0.0429	0.5234	0.0247	0.0576	1.3925
نصر للتعمير والإسكان	0.2413	0.0579	0.0671	1.1946	0.0332	0.0532	2.0438
هيرمس القابضة	0.0990	0.0226	0.0747	0.4163	0.0135	0.0766	0.6846
أجواء للصناعات الغذائية	0.1998	0.0534	0.0718	0.6959	0.0327	0.1120	1.1786
البنك التجارى الدولى	0.1028	0.0302	0.0551	0.5160	0.0209	0.0599	1.0867

ويتضح من الجدول السابق أن التجميع على مستوى الزمن وباستخدام المجموع الكلى كأسلوب للتجميع يؤدي إلى انخفاض دقة تنبؤ النموذج. أما باستخدام مقاييس النزعة المركزية فى تجميع البيانات على مستوى الزمن فإن دقة تنبؤ النموذج التجميعي تزداد.

(2-2-4) أثر مشكلة التجميع على مستوى المؤسسات

أولاً: القوة التفسيرية للنموذج

للتعرف على أثر التجميع على مستوى المؤسسات على القوة التفسيرية فى نموذج الانحدار الخطى المتعدد، تم الإستعانة ببيانات 14 قطاعاً إقتصادياً مختلفاً، وقياس قيم معامل التحديد قبل وبعد تجميع البيانات على مستوى المؤسسات "الشركات والبنوك". ويُلخص جدول (7) أهم تلك النتائج .

جدول (7)

معامل تحديد نموذج الانحدار الخطى المتعدد R^2 فى ظل التجميع على مستوى المؤسسات

نوع التجميع			تجميع البيانات على مستوى المؤسسات	
أسلوب التجميع القطاعات	النموذج التفصيلي	الوسط الحسابي	الوسط الحسابي	المجموع الكلى
البنوك	95.8%	96.4%	91%	97.1%
الموارد الأساسية	93.6%	85.9%	99.4%	85.4%
الاتصالات	97.5%	83.3%	97.7%	83.9%
الأغذية والمشروبات	91.3%	59.2%	59.5%	97.3%
التشيد والبناء	94.1%	59.4%	59.4%	90.5%
الخدمات المالية	99.8%	95.8%	80.8%	99.9%
السياحة والترفيه	98.4%	99.2%	97%	99.1%
العقارات	98.2%	98.9%	96.7%	98.9%
الغاز والبترو	99.6%	98.8%	98.8%	97.2%
الكيموايات	89.3%	75.5%	81.2%	64.8%
خدمات ومنتجات صناعية	97.0%	96.6%	98.8%	96.3%
الرعاية الصحية والأدوية	83.2%	78%	78.5%	92.1%
منتجات منزلية وشخصية	97.2%	91.9%	98.0%	92.0%
الموزعين وتجار تجزئة	77.4%	84%	84%	85.9%

ويتضح من الجدول السابق مدى تأثير القوة التفسيرية لنموذج الانحدار الخطى بمشكلة التجميع على مستوى المؤسسات وفيما يلي استعراض لأهم تلك النتائج:

- بتطبيق التجميع على مستوى المؤسسات باستخدام المجموع الكلى، فإن قيم معامل التحديد تتفاوت ما بين الزيادة والنقصان دون الوصول إلى أثر محدد. حيث يزداد معامل التحديد فى القطاعات الآتية: قطاع البنوك، وقطاع الأغذية والمشروبات، وقطاع الخدمات المالية باستثناء البنوك، وقطاع العقارات، وقطاع الرعاية الصحية والأدوية، وقطاع الموزعين وتجار التجزئة.
- وباستخدام أحد مقاييس النزعة المركزية كالوسط الحسابى أو الوسيط فى تطبيق ذلك النوع من التجميع، فإن قيم معامل التحديد تنخفض فى النموذج التجميعى وذلك فى جميع القطاعات فيما عدا:
 - قطاع البنوك وقطاع السياحة والترفيه وقطاع العقارات وقطاع الموزعين وتجار التجزئة وذلك فى حالة استخدام الوسط الحسابى.
 - أما باستخدام الوسيط فى تجميع البيانات على مستوى المؤسسات فإن قيم معامل التحديد تنخفض فى النموذج التجميعى وذلك فى جميع القطاعات فيما عدا: قطاع الموارد الأساسية وقطاع الاتصالات وقطاع الخدمات والمنتجات الصناعية وقطاع المنتجات المنزلية وقطاع الموزعين وتجار التجزئة.

ثانياً: كفاءة مقدرات معالم النموذج

لتلخيص والتعرف على مدى تأثر الخطأ المعيارى لمقدرات معالم النموذج بمشكلة التجميع على مستوى المؤسسات. فقد تم تجميع بيانات عدد 14 قطاعاً إقتصادياً فى نموذج الانحدار الخطى المتعدد. وجدول (8) التالى يلخص ذلك الأثر فى نموذج الانحدار الخطى.

جدول (8)

الخطأ المعيارى لمقدرات نموذج الانحدار الخطى المتعدد فى ظل التجميع على مستوى المؤسسات

تجميع البيانات شهرياً			نوع التجميع		
المجموع الكلى	الوسيط	الوسط الحسابى	أسلوب التجميع	نموذج الانحدار الخطى المقدر	القطاعات
			النموذج التفصيلى		
-1.0310	0.0134	-0.0162	-0.0627	قيمة المقدر (β_0)	البنوك
0.7327	0.0878	0.0703	0.0279	الخطأ المعيارى للمقدر (β_0)	
-0.3489	-0.2603	-0.4029	0.0953	قيمة المقدر (β_1)	
0.0762	0.0914	0.0793	0.0279	الخطأ المعيارى للمقدر (β_1)	
1.3097	1.2017	1.3569	0.8782	قيمة المقدر (β_2)	
0.0736	0.0879	0.0767	0.0271	الخطأ المعيارى للمقدر (β_2)	
-104.21	-0.4108	-15.490	-3.0187	قيمة المقدر (β_0)	الموارد الأساسية
9.256	0.0448	1.349	0.6678	الخطأ المعيارى للمقدر (β_0)	
-0.7715	0.3648	-0.8543	-1.2066	قيمة المقدر (β_1)	
0.1903	0.0209	0.1927	0.0803	الخطأ المعيارى للمقدر (β_1)	
1.7890	0.6329	1.8694	2.0144	قيمة المقدر (β_2)	
0.1837	0.0194	0.1864	0.0783	الخطأ المعيارى للمقدر (β_2)	
-11.66	-0.2564	-3.178	-0.0171	قيمة المقدر (β_0)	الاتصالات
1.891	0.0463	0.4927	0.1468	الخطأ المعيارى للمقدر (β_0)	
0.1983	0.4492	0.1943	0.2259	قيمة المقدر (β_1)	
0.1181	0.1157	0.1181	0.0618	الخطأ المعيارى للمقدر (β_1)	
0.8384	0.5560	0.8472	0.7498	قيمة المقدر (β_2)	
0.1156	0.1120	0.1157	0.0601	الخطأ المعيارى للمقدر (β_2)	

جدول (8)

الخطأ المعيارى لمقدرات نموذج الانحدار الخطى المتعدد فى ظل التجميع على مستوى المؤسسات

تجميع البيانات شهرياً			نوع التجميع		
المجموع الكلى	الوسيط	الوسط الحسابى	أسلوب التجميع	نموذج الانحدار الخطى المقدر	القطاعات
			النموذج التفصيلى		
-14.206	3.335	4.207	0.5019	قيمة المقدر (β_0)	الأغذية والمشروبات
1.401	0.2413	0.2708	0.0355	الخطأ المعياري للمقدر (β_0)	
0.1789	-2.002	-2.632	-0.3015	قيمة المقدر (β_1)	
0.0549	0.1689	0.1695	0.0225	الخطأ المعياري للمقدر (β_1)	
0.8273	2.629	3.224	1.2139	قيمة المقدر (β_2)	
0.0505	0.1647	0.1638	0.0214	الخطأ المعياري للمقدر (β_2)	
-34.77	2.2328	-0.7286	-0.1361	قيمة المقدر (β_0)	التشييد والبناء
4.487	0.3005	0.6452	0.9182	الخطأ المعياري للمقدر (β_0)	
-1.0113	-1.845	-2.740	-0.9183	قيمة المقدر (β_1)	
0.1437	0.1799	0.221	0.0449	الخطأ المعياري للمقدر (β_1)	
1.980	2.566	3.603	1.8341	قيمة المقدر (β_2)	
0.1437	0.1764	0.2141	0.0437	الخطأ المعياري للمقدر (β_2)	
-0.5575	0.1372	0.0106	-0.0258	قيمة المقدر (β_0)	الخدمات المالية
0.0528	0.0441	0.0375	0.0043	الخطأ المعياري للمقدر (β_0)	
0.5178	-0.8124	-0.0627	0.4499	قيمة المقدر (β_1)	
0.0164	0.1280	0.0794	0.007	الخطأ المعياري للمقدر (β_1)	
0.4898	1.693	1.0206	0.5472	قيمة المقدر (β_2)	
0.0152	0.1227	0.0733	0.0066	الخطأ المعياري للمقدر (β_2)	
-2.129	-0.5932	-0.4283	-0.2021	قيمة المقدر (β_0)	السياحة والترفيه
0.1283	0.0408	0.0233	0.0172	الخطأ المعياري للمقدر (β_0)	
0.1787	0.2409	0.1956	0.1809	قيمة المقدر (β_1)	
0.0499	0.0870	0.0463	0.0233	الخطأ المعياري للمقدر (β_1)	
0.8267	0.7792	0.8147	0.8041	قيمة المقدر (β_2)	
0.0476	0.0836	0.0442	0.0221	الخطأ المعياري للمقدر (β_2)	
-3.2007	-0.0751	-0.2889	-0.0858	قيمة المقدر (β_0)	العقارات
0.7225	0.0497	0.0369	0.0133	الخطأ المعياري للمقدر (β_0)	
0.1541	0.2957	0.2089	0.2797	قيمة المقدر (β_1)	
0.0329	0.0361	0.0309	0.0113	الخطأ المعياري للمقدر (β_1)	
0.8263	0.6854	0.7836	0.7019	قيمة المقدر (β_2)	
0.0309	0.0342	0.0290	0.0106	الخطأ المعياري للمقدر (β_2)	
0.0621	-0.0781	-0.0781	0.0064	قيمة المقدر (β_0)	الغاز والبتروك
0.2410	0.0904	0.0904	0.0474	الخطأ المعياري للمقدر (β_0)	
0.4261	0.3523	0.3523	0.4062	قيمة المقدر (β_1)	
0.0577	0.0443	0.0443	0.0417	الخطأ المعياري للمقدر (β_1)	
0.5693	0.6452	0.6452	0.5897	قيمة المقدر (β_2)	
0.0573	0.0435	0.0435	0.0409	الخطأ المعياري للمقدر (β_2)	
13.233	1.4810	3.1687	1.0453	قيمة المقدر (β_0)	الكيمائيات
2.629	0.3604	0.5225	0.1550	الخطأ المعياري للمقدر (β_0)	
-1.2161	-1.1044	-1.3086	-0.9299	قيمة المقدر (β_1)	
0.1870	0.3656	0.3403	0.0929	الخطأ المعياري للمقدر (β_1)	
2.055	1.9286	2.1193	1.8187	قيمة المقدر (β_2)	

تابع جدول (8)

الخطأ المعياري لمقدرات نموذج الانحدار الخطى المتعدد فى ظل التجميع على مستوى المؤسسات

جامعة القاهرة - معهد الدراسات والبحوث الإحصائية

0.1817	0.3625	0.3378	0.0913	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	الخدمات والمنتجات الصناعية
-0.529	-0.1107	-0.1068	0.2224	قيمة المقدر ($\hat{\beta}_0$)	
1.129	0.0271	0.1060	0.0486	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
-0.1411	0.5144	-0.0654	-0.1666	قيمة المقدر ($\hat{\beta}_1$)	
0.0721	0.0217	0.0689	0.0253	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
1.0957	0.4883	1.0292	1.0596	قيمة المقدر ($\hat{\beta}_2$)	
0.0672	0.0199	0.0641	0.0239	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	الرعاية الصحية والأدوية
-8.392	-1.0033	-2.1419	-0.7691	قيمة المقدر ($\hat{\beta}_0$)	
1.198	0.3503	0.4124	0.1611	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
-0.1310	-0.1586	-0.7054	-0.0124	قيمة المقدر ($\hat{\beta}_1$)	
0.1067	0.1028	0.1407	0.0497	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
1.1197	1.1256	1.6858	0.9766	قيمة المقدر ($\hat{\beta}_2$)	
0.1016	0.1028	0.1359	0.0483	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	المنتجات المنزلية
-3.017	-0.0619	-0.4646	-0.0594	قيمة المقدر ($\hat{\beta}_0$)	
1.033	0.0160	0.1507	0.0730	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
0.0493	0.4685	0.0485	0.0510	قيمة المقدر ($\hat{\beta}_1$)	
0.0935	0.0394	0.0927	0.0368	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
0.9388	0.5356	0.9410	0.9186	قيمة المقدر ($\hat{\beta}_2$)	
0.0911	0.0377	0.0903	0.0359	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	الموزعين وتجارة التجزئة
0.1695	0.0159	0.0159	0.1224	قيمة المقدر ($\hat{\beta}_0$)	
0.1035	0.0621	0.0621	0.0580	الخطأ المعياري للمقدر ($\hat{\beta}_0$)	
-0.7981	-0.6918	-0.6918	1.0948	قيمة المقدر ($\hat{\beta}_1$)	
0.1609	0.1441	0.1441	0.1178	الخطأ المعياري للمقدر ($\hat{\beta}_1$)	
1.7077	1.6199	1.6199	1.9908	قيمة المقدر ($\hat{\beta}_2$)	
0.1573	0.1403	0.1403	0.1139	الخطأ المعياري للمقدر ($\hat{\beta}_2$)	

ومن الجدول السابق تُعرض أهم النتائج كالآتي:

- يزداد الخطأ المعياري لمقدرات معالم النموذج التجميعي إذا تم التجميع بأسلوب المجموع الكلي أو الوسط الحسابي، وذلك بالنسبة لمقدرات معالم الحد الثابت و مقدرات معالم الميل.
- تتفاوت قيم الخطأ المعياري لمقدرات معالم نموذج الانحدار الخطي التجميعي الذي يعتمد على الوسيط في تكوين متغيراته، وذلك حيث يزداد الخطأ المعياري في 12 قطاعاً إقتصادياً ويقل في قطاع الموارد الأساسية، وقطاع الخدمات والمنتجات الصناعية.

ثالثاً: القدرة التنبؤية لنموذج الانحدار الخطي

يُلخص جدول (9) كيفية تأثير القدرة التنبؤية لنموذج الانحدار الخطي المقدر بمشكلة تجميع البيانات على مستوى المؤسسات، وذلك باستخدام بيانات يومية عن الفترة من أول يناير 2007 م إلى آخر يناير 2015م لعدد 14 قطاعاً إقتصادياً. وباختبار مدى استقرار التقديرات والذي يعتمد على إضافة بيانات جديدة من أول فبراير 2015م إلى آخر أغسطس 2015م. ثم اختبار الفرق بين متوسط القيم الفعلية بعد زيادة البيانات الجديدة μ_{Y_2} ، ومتوسط القيم الفعلية قبل زيادة البيانات الجديدة μ_{Y_1} باستخدام اختبار t للفرق بين متوسطي مجتمعين مستقلين وذلك بنسبة خطأ مسموح بها تساوي 0.05.

جدول (9)

اختبار t لدراسة القدرة التنبؤية لنموذج الانحدار الخطي في ظل لتجميع على مستوى الزمن

جامعة القاهرة – معهد الدراسات والبحوث الإحصائية

جميع على مستوى المؤسسات			نوع التجميع		
المجموع الكلي	الوسيط	الوسط الحسابي	النموذج التفصيلي	معايير الاختبار	أسلوب التجميع القطاعات
1.19	0.002	0.191	0.174	الفرق بين المتوسطين	البنوك
0.63	0.01	1.25	1.43	قيمة "ت" المحسوبة	
0.528*	0.99*	0.211*	0.152* ³	القيمة الاحتمالية	
-41.3	-0.797	-6.10	-5.64	الفرق بين المتوسطين	الموارد الأساسية
-2.83	-1.91	-2.75	-1.69	قيمة "ت" المحسوبة	
0.005	0.056*	0.006	0.09*	القيمة الاحتمالية	
-3.08	-0.335	-0.811	-0.81	الفرق بين المتوسطين	الاتصالات
-1.84	-1.5	-1.97	-0.75	قيمة "ت" المحسوبة	
0.065*	0.133*	0.049*	0.452*	القيمة الاحتمالية	
8.10	-0.894	-0.655	-0.386	الفرق بين المتوسطين	الأغذية و المشروبات
2.27	-4.44	-3.31	-3.7	قيمة "ت" المحسوبة	
0.023	0.0001	0.001	0.0001	القيمة الاحتمالية	
10.26	-1.246	-0.706	-0.820	الفرق بين المتوسطين	التشيد والبناء
1.70	-5.76	-2.45	-1.95	قيمة "ت" المحسوبة	
0.090*	0.0001	0.014	0.051*	القيمة الاحتمالية	
1.361	-0.188	-0.107	-0.143	الفرق بين المتوسطين	الخدمات المالية باستثناء البنوك
1.50	-3.62	-0.97	-1.44	قيمة "ت" المحسوبة	
0.133*	0.0001	0.333*	0.151*	القيمة الاحتمالية	
0.807	-4.193	-2.976	-0.371	الفرق بين المتوسطين	السياحة والترفيه
0.84	-31.08	-19.36	-2.67	قيمة "ت" المحسوبة	
0.400*	0.0001	0.0001	0.007	القيمة الاحتمالية	
-2.01	-0.071	-0.110	-0.113	الفرق بين المتوسطين	العقارات
-0.76	-0.74	-0.89	-1.22	قيمة "ت" المحسوبة	
0.448*	0.462*	0.371*	0.221*	القيمة الاحتمالية	
-0.782	-0.647	-0.647	-0.577	الفرق بين المتوسطين	الغاز والبتترول
-1.59	-1.74	-1.74	-0.86	قيمة "ت" المحسوبة	
0.112*	0.082*	0.082*	0.391*	القيمة الاحتمالية	
-0.95	-6.121	-5.352	-1.177	الفرق بين المتوسطين	الكيمويات
-0.58	-9.46	-8.54	-2.31	قيمة "ت" المحسوبة	
0.562*	0.0001	0.0001	0.021	القيمة الاحتمالية	
4.55	-0.223	0.349	0.375	الفرق بين المتوسطين	الخدمات والمنتجات الصناعية
1.42	-1.96	1.11	1.09	قيمة "ت" المحسوبة	
0.154*	0.050*	0.265*	0.275*	القيمة الاحتمالية	
2.61	-1.220	-0.805	-0.607	الفرق بين المتوسطين	الرعاية الصحية والأدوية
1.26	-3.78	-2.63	-2.45	قيمة "ت" المحسوبة	
0.207*	0.0001	0.009	0.014	القيمة الاحتمالية	
4.52	-0.100	0.627	0.637	الفرق بين المتوسطين	المنتجات المنزلية والشخصية
2.00	-2.01	1.94	1.13	قيمة "ت" المحسوبة	
0.045	0.044	0.053*	0.258*	القيمة الاحتمالية	
0.054	0.0562	0.0562	0.0751	الفرق بين المتوسطين	الموزعين وتجار التجزئة
0.40	0.87	0.87	1.21	قيمة "ت" المحسوبة	
0.692*	0.382*	0.382*	0.226*	القيمة الاحتمالية	

ومن الجدول السابق يمكن الحصول على النتائج التالية:

³ يأخذ القرار بقبول أو رفض فرض العدم وفقاً للقيمة الاحتمالية p-value وذلك بنسبة خطأ مسموح به مقدار 0.05. كما يعبر الرمز "*" على قبول فرض العدم وتحقيق استقرار تقديرات معالم نموذج الانحدار.

- إذا كان النموذج التفصيلى له قدرة تنبؤية فإن التجميع على مستوى المؤسسات يحافظ على تلك القدرة التنبؤية، واستقرار تقديرات معالم النموذج الخطى فى بعض القطاعات. وذلك كما فى قطاع البنوك، وقطاع الاتصالات، وقطاع الخدمات المالية باستثناء البنوك، وقطاع العقارات، وقطاع الغاز والبترو، وقطاع الخدمات والمنتجات الصناعية، وقطاع الموزعين وتجار التجزئة.
- بينما تقل القدرة التنبؤية بسبب مشكلة تجميع البيانات على مستوى المؤسسات وذلك كما فى قطاع الموارد الأساسية، وقطاع التشيد والبناء، وقطاع المنتجات المنزلية والشخصية.
- إذا كان النموذج التفصيلى ليس لديه القدرة على التنبؤ فإن تجميع البيانات على مستوى المؤسسات لا ينعكس استقرار تقديرات معالم النموذج وانعكاسها على قدرته التنبؤية. وذلك كما فى القطاعات الآتية: قطاع الأغذية والمشروبات، وقطاع السياحة والترفيه، وقطاع الكيماويات، وقطاع الرعاية الصحية والأدوية. وإذا حدث تحسن فى قدرة النموذج التجميعى على التنبؤ فإنه ذلك لا يحدث إلا فى ظل التجميع باستخدام المجموع الكلى.

رابعاً: دقة تنبؤ نموذج الانحدار الخطى

وبالتجميع على مستوى المؤسسات، فقد تم تلخيص أهم نتائج جذر متوسط مربعات الخطأ فى نموذج الانحدار الخطى المتعدد كما فى جدول (10) التالى.

جدول (10)

جذر متوسط مربعات الخطأ RMSE لنموذج الانحدار الخطى المتعدد فى ظل التجميع على مستوى المؤسسات

تجميع البيانات على مستوى المؤسسات			نوع التجميع	
المجموع الكلى	الوسيط	الوسط الحسابى	النموذج التفصيلى	أسلوب التجميع القطاعات
10.201	1.2126	0.91916	2.6259	البنوك
174.76	1.0682	26.060	69.527	الموارد الأساسية
21.035	1.0884	5.2935	10.713	الاتصالات
20.085	2.0486	4.7449	4.1563	الأغذية والمشروبات
63.467	4.8669	5.9622	13.586	التشيد والبناء
1.1439	0.7647	0.7285	0.4255	الخدمات المالية
2.9931	0.9929	0.5277	1.2784	السياحة والترفيه
8.922	0.56212	0.4205	1.7429	العقارات
2.6228	1.3172	1.3172	1.9225	الغاز والبترو
30.282	12.543	13.364	12.828	الكيماويات
19.298	0.39349	1.8223	5.8647	الخدمات الصناعية
19.019	5.1700	5.0237	7.3069	الرعاية الصحية والأدوية
19.876	0.22639	2.8636	7.6278	منتجات منزلية وشخصية
1.6005	0.80607	0.80607	1.2099	الموزعين وتجار التجزئة

ومن الجدول السابق يمكن استنتاج الآتى:

- يؤدى التجميع على مستوى المؤسسات إلى زيادة قيم RMSE فى النموذج التجميعى، والذي يعتمد على المجموع الكلى فى تكوين المتغيرات التجميعية. مما يؤكد على انخفاض دقة التنبؤ للنموذج بسبب التجميع على مستوى المؤسسات.

- إذا تم استخدام أحد مقاييس النزعة المركزية للتجميع على مستوى المؤسسات كالوسط الحسابى أو الوسيط، فإن قيمة جذر متوسط مربعات الخطأ تقل فى النموذج التجميعى بالمقارنة بقيمته فى النموذج التفصيلى مما يدل على ارتفاع دقة تنبؤ النموذج بالتجميع.

وذلك كما فى جميع القطاعات 14 محل الدراسة، ولكن فيما عدا قطاع الأغذية والمشروبات، وقطاع الخدمات المالية باستثناء البنوك، وقطاع الكيماويات حيث تقل دقة تنبؤ النموذج التجميعى الذى يعتمد على مقاييس النزعة المركزية مثل: الوسط الحسابى أو الوسيط.

5- الخلاصة

تؤدى مشكلة التجميع إلى العديد من الآثار على جودة توفيق نموذج الانحدار، وقد تم قياس ذلك من خلال القوة التفسيرية لنموذج الانحدار وقد تم قياس معامل التحديد كمقياس لها. كما تم قياس الخطأ المعياري لمقدرات معالم النموذج وذلك لدراسة أثر مشكلة التجميع على كفاءة تلك مقدرات. وللتعرف على مدى تأثير القدرة التنبؤية لنموذج الانحدار الخطى، تم استخدام اختبار مدى استقرار التقديرات لنموذج الانحدار الخطى. وأخيراً لقياس أثر مشكلة التجميع على دقة تنبؤات نموذج الانحدار الخطى، تم الاستعانة بجذر متوسط مربعات الخطأ كمقياس لها. وذلك بتجميع البيانات على مستوى الزمن ومستوى المؤسسات محل الدراسة وباستخدام المجموع الكلى ومقاييس النزعة المركزية مثل: الوسط الحسابى والوسيط كأساليب لتجميع البيانات. وفى ظل البيانات محل الدراسة تتلخص أهم نتائج الدراسة القياسية فيما يلى:

- يتوقف أثر مشكلة التجميع على القوة التفسيرية على نوع التجميع حيث تؤدى تلك المشكلة إلى زيادة قيم معامل التحديد إذا تم تجميع البيانات على مستوى الزمن. ولكن لم يتحدد أثر واضح للتجميع على قيم معامل التحديد إذا تم التجميع على مستوى الوحدات القطاعية.
- تؤدى مشكلة التجميع إلى انخفاض كفاءة مقدرات ols بغض النظر عن أسلوب أو نوع التجميع.
- يتوقف أثر مشكلة التجميع على القدرة التنبؤية للنموذج على نوع التجميع حيث يُحسن التجميع على مستوى الزمن من قدرة نموذج الانحدار الخطى على التنبؤ، بينما يضعف التجميع على مستوى المؤسسات من قدرة النموذج على التنبؤ.
- يتوقف أثر مشكلة التجميع على دقة التنبؤات على أسلوب التجميع حيث تنخفض دقة تنبؤات النموذج التجميعى الذى يعتمد على المجموع الكلى فى التجميع و زيادة دقة التنبؤات التى تعتمد على أحد مقاييس النزعة المركزية فى التجميع مثل الوسط الحسابى والوسيط.

6- المراجع

- [1] Garderen, K.J.V., Lee, K. and Pesaran, M.H. (2000). " Cross Section Aggregation of Non-Linear Models." Journal of Econometrics ,95,pp.285-331.
- [2] Gordevicius, J., Gamper, J. and Bohlen, M. (2012). "Parsimonious Temporal Aggregation." The VLDB Journal, 21, pp.309-332.
- [3] Hassler, U. (2014). " Persistence under Temporal Aggregation and differencing. " Economics Letters,124,pp.318-322.

- [4] Kabacoff, R.I. (2011). "R in Action Data Analysis and Graphics with R." Shelter Island.
- [5] Lutero, G. (2010). " The Aggregation Problem in its Historical Perspective: A Summary overview ." Methods Development of quarterly national accounts. ISTAT, National Accounts Directorate.
- [6] Monteforte, L. (2006). "Aggregation Bias in Macro Models : Does it matter for the Euro area? " Economic Modelling , 24,pp.236-261.
- [7] Petkovic, A. and Veredas, D. (2009). " Aggregation of Linear Models for Panel Data." International Association For Research and Teaching, working paper no .12.
- [8] Plassmann, E. (2002). "Econometric Modelling of European Money Demand." Physica -Verlag Berlin Heidelberg New York, pp.81-88.
- [9] Rossana , R.J. and Seater , J.J. (1995). " Temporal Aggregation and Economic Time Series." Journal of Business & Economic Statistics , 13, pp. 441-451.
- [10] Theil, H. (1954). " Linear Aggregation of Economic relationships." Amsterdam: North Holland.
- [11] Yager,R.,R.(1998). " Fusion of ordinal Information using Weighted Median Aggregation." International Journal of Approximate Reasoning 18,pp.35-52.

اهمية قواعد وإجراءات السلامة والصحة المهنية علي مراحل إدارة الأزمات والكوارث

(دراسة تحليلية على إحدى شركات الإتصالات)

عصام عطية عبدالمنعم

الملخص

تعد إدارة السلامة والصحة المهنية وفي ظل العديد من الازمات والكوارث في الالونة الاخيرة من المداخل العلمية الحديثة التي تمارسها المنظمات الانتاجية والخدمية على حد سواء لتوفير بيئة عمل آمنة من المخاطر المختلفة ورفع مستوى كفاءة وسائل الوقاية، ويهدف البحث الى تحليل العلاقة بين ادارة السلامة والصحة المهنية والازمات والكوارث ودور إدارة السلامة والصحة المهنية في الحد من الأزمات والكوارث وكلما استطاعت الشركة المصرية للاتصالات الساعية الى الاستمرار والنمو في ظل العديد من الازمات والكوارث في الالونة الاخيرة وفي ظل المنافسة المتزايدة من تطبيق قواعد وإجراءات السلامة والصحة المهنية كلما أدى ذلك الى مواجهة الازمات والكوارث التي قد تجابهها بالاسلوب الذي يمكنها من معالجتها وأيجاد السبل الكفيلة لتفاديها مستقبلاً عند تكرارها خاصة وان الظروف الاقتصادية التي تمر بها الشركة المصرية للاتصالات تتطلب المزيد من التحسين والانتباه الى مصادر الازمات والكوارث وادارتها بالاسلوب العلمي المناسب لتجاوزها دون اضرار ناهيك عن الخسائر المادية والمعنوية التي قد تتعرض لها الشركة جراء هذه الازمات . وقد توصلت الدراسة الى عدة استنتاجات اعتمدت عليها التوصيات المناسبة .

الكلمات الدالة

- عينة طبقية
- معاملات الارتباطات البسيطة
- معاملات الانحدار القياسية

طالب دكتوراة جودة في معهد الدراسات والبحوث الاحصائية

مقدمة البحث

لقد أدى التطور التقني الذي شهده العالم إلى ظهور العديد من المخاطر والازمات التي ينبغي على الإنسان إدراكها وتجنب الوقوع في مسبباتها ، فأماكن العمل المتعددة والمختلفة من ورش ومصانع ومختبرات ومعامل تعتبر بيئات عمل تكثر فيها العديد من المخاطر المهنية التي من الممكن ان تتسبب في حدوث العديد من الازمات والكوارث في بيئة العمل. (موقع دليل السلامة والصحة المهنية – <http://www.education.gov.bh>)

لذلك فإن توفير بيئة عمل آمنة من مخاطر الصناعات المختلفة ورفع مستوى كفاءة وسائل الوقاية سيؤدي بلا شك إلى الحد من الإصابات والأمراض المهنية ، وحماية العاملين من الحوادث والتي قد تتسبب في حدوث العديد من الازمات والكوارث .

(المركز العالمي للصحة والسلامة المهنية - <http://www.salama-libya.org>)

وفي ظل التأثير السلبي الذي تخلفه الأمراض والإصابات المرتبطة بالعمل على صحة العمال وإنتاجيتهم وبالتالي على عائلاتهم والوضع الاقتصادي والرفاهية الاجتماعية في البلاد، ازدادت التوعية حول الصحة والسلامة المهنية والازمات والكوارث في العالم بأكمله. (خوان سومافيا ، 2005).

وما زالت الأمراض والحوادث المهنية أهم أسباب الإصابات والوفيات للعاملين في بيئة العمل ويمكن تفادي أغلب الحوادث والازمات والكوارث عبر "وقاية سليمة تدعمها ممارسات ملائمة للتفتيش والتقرير وترشدها اتفاقيات منظمة العمل الدولية". (منظمة العمل الدولية ، جنيف، 2007)

ومن هنا زاد الاهتمام بالسلامة والصحة المهنية وسنت القوانين والتشريعات وانهضت المؤتمرات المحلية والدولية المتلاحقة الهادفة لحماية الإنسان في العمل مثل الاتفاقية الدولية (اتفاقية السلامة والصحة والمهنية) رقم (155) لعام 1981 والتي تهدف إلى الوقاية من الحوادث والإصابات الصحية الناجمة عن العمل والتقليل من المخاطر المرتبطة ببيئة العمل بالإضافة إلى قانون العمل المصري رقم (12) الصادر لسنة (2003). ومن أجل تمكين الشركات من تحقيق أهدافها والقيام بدورها لابد من توفير إجراءات السلامة في الشركات وإيجاد بيئة عمل خالية من المخاطر والازمات وإلى تعريف العاملين بقواعد السلامة بهدف الوقاية من مخاطر العمل وتحقيق أكبر قدر من الصحة المهنية وأقل قدر من الخسائر المادية والبشرية وهذا يساعد على زيادة الإنتاج ويخفض التكلفة (أبو عبدون، 2003).

لذلك فإن هذه الدراسة سوف تسلط الضوء على دور إدارة السلامة والصحة المهنية في الحد من الازمات والكوارث

مشكلة البحث

تعتبر المصرية للاتصالات شركة رائدة في مجال الاتصالات السلكية واللاسلكية ونقل المعلومات ذات السرعات الفائقة وهي اعرق شركات الاتصالات في افريقيا والشرق الاوسط .
إلا إن القصور في تطبيق اجراءات ومتطلبات السلامة والصحة المهنية في هذه الشركة وعدم مطابقتها بالمعايير الدولية الموضوعة من قبل منظمة العمل الدولية (ILO) وإدارة السلامة والصحة المهنية (OSHA) ومنظمة الصحة العالمية (WHO) يؤدي ذلك إلى وقوع بعض الحوادث والازمات وقد تشمل الشركة نفسها أو قد تمتد إلى خارجها (المديفر، 2005) لذا يجب أن يتوفر الجهد لرفع مستوى الوعي لدى كل من العاملين والادارة بأهمية السلامة والصحة المهنية في مواقع العمل بالشركة، كما لا بد من التركيز على مجال التدريب والاستفادة من برامج السلامة الشاملة (كوهن وآخرون، 1995) ،

مما تقدم فان مشكلة البحث تبرز من خلال ما يلي :-

- هل هناك ارتباط بين عناصر قواعد واجراءات السلامة والصحة المهنية مع عناصر مراحل إدارة الازمات والكوارث في الشركة المصرية للاتصالات ؟
- ترتيب عناصر قواعد واجراءات السلامة والصحة المهنية وفقا للعلاقة مع عناصر مراحل إدارة الازمات والكوارث في الشركة المصرية للاتصالات .

أهمية البحث

- إنبثقت أهمية الدراسة من تحديد الآثار الإيجابية من تطبيق ادارة السلامة والصحة المهنية في الحد من الأزمات و الكوارث من خلال تطبيق منظومة ادارة السلامة والصحة المهنية بكفاءة و فاعلية مما تساعد في مقاومة الأزمات و الكوارث و تعجل من سرعة إحتوائها و إعادة الأمن و الإستقرار و إزالة آثارها السلبية و كذلك دراسة الأزمات و الكوارث السابقة بما يكفل فاعلية مواجهتها و إحباط آثارها مستقبلاً ،
- وإن عملية تقدير القيمة الاقتصادية للعنصر البشري في أي وحدة إنتاجية يجب أن تنال اهتماماً كبيراً وأن يكون هناك أسلوب دقيق يجب إتباعه واتساع دائرة تطبيقه في جميع الشركات للخروج بصورة دقيقة عن قيمة الثروة البشرية ومدى العائد من استثماره في التدريب والتعليم والخبرة حيث توضح أهمية الدراسة فيما يلي :
- تساهم هذه الدراسة للباحث في إثراء معلوماته حول السلامة والصحة المهنية ، وكيفية تجنب المخاطر المهنية والازمات والوقاية منها .

- تسهم الدراسة في معرفة العاملين بإشترطات السلامة والصحة المهنية عند أدائهم لأعمالهم وبالتالي التقليل من حجم الخسائر سواء المادية أو البشرية في الشركة المصرية للاتصالات
- تساعد هذه الدراسة العاملين بالشركات في الأخذ بكافة احتياطات السلامة المهنية والقواعد والإجراءات الواجب إتباعها في بيئة العمل ، وذلك من أجل خلق جو آمن وبيئة خالية من المخاطر والازمات في جميع المجالات.
- توضح الدراسة مدى احتياج إدارة الشركة المصرية للاتصالات لنتائج هذه الدراسة والتي من شأنها القيام بتطوير قواعد وإجراءات السلامة والصحة المهنية واساليب إدارة الازمات والكوارث .
- تساعد نتائج هذه الدراسة المسؤولين عن الشركة المصرية للاتصالات في تفعيل قواعد وإجراءات السلامة والصحة المهنية عند قيامهم بعملهم في الشركة.

أهداف البحث

أن الهدف الرئيسي من هذا البحث هو ابراز اهمية دور تطبيق قواعد واجراءات السلامة والصحة المهنية بالشركة المصرية للاتصالات في الحد من الازمات والكوارث وذلك من خلال قياس المتغيرات المستقلة المتمثلة في قواعد واجراءات السلامة والصحة المهنية ، والمتغيرات التابعة المتمثلة في مراحل إدارة الازمة أو الكارثة ، وذلك ببعض العبارات التي تناسب كل متغير.

منهج البحث وخطواته

سوف يتم إستخدام المنهج الوصفي التحليلي عن طريق مدخل المسح بإستخدام الإستبانة كأداة لجمع البيانات بالإضافة الى اساليب اخرى وفيما يلي توضيحها .

جمع البيانات

سوف يعتمد اسلوب العمل في أنجاز الجانب النظري من الدراسة الأسلوب الوصفي باستخدام المراجع العلمية والدوريات وأدبيات إدارة الإنتاج والعمليات فضلا عن شبكة الانترنت ومحتوياتها الحديثة. اما الجانب الميداني من الدراسة سوف يعتمد بشكل رئيسي على استمارات الاستبيان التي جاءت متوافقة ومستمدة من الجانب النظري من الدراسة بما يمكن تحديد قواعد واجراءات السلامة والصحة المهنية ومراحل إدارة الازمات والكوارث بالشركة المصرية للاتصالات اذ وزعت (275) مائتان وخمسة وسبعون استمارة على مسؤولي الوحدات ورؤساء الاقسام زيادة على ذلك المقابلات الشخصية والزيارات الميدانية للادارات ومواقع الشركة المختلفة بالمحافظات المختلفة . كما اعتمد اسلوب ليكرت الخماسي على تحليل الاستمارة وفضلاً عن الارتباط بين متغيرات الدراسة الى جانب

التكرارات والنسب المئوية والمتوسطات والانحرافات المعيارية ومعامل الاختلاف لوصف وتشخيص متغيرات الدراسة

مجتمع الدراسة :

تقوم الشركة المصرية للاتصالات بدور حيوى وهام فى خدمة المجتمع من خلال توفير خدمة الانترنت فائق السرعة بما ساهم فى سهولة تدوال البيانات بما يرفع من الدخل القوى للبلاد .

وقد قام الباحث بسحب عينة طبقية من مختلف المستويات الوظيفية بالادارة العامة للسلامة والصحة المهنية وجميع الاقسام التابعة لها فى قطاعات الشركة المختلفة بالمحافظات وكذلك الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى قطاعات الشركة المختلفة بالمحافظات ومبنى سنترال القبة وحيث أن مجتمع الدراسة المتمثل فى الادارة العامة للسلامة والصحة المهنية وجميع الاقسام التابعة لها فى قطاعات الشركة المصرية للاتصالات المختلفة بالمحافظات والبالغ عدد مفردات الدراسة من واقع سجلات شئون العاملين وهم (90) مفردة وكذلك الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى قطاعات الشركة المصرية للاتصالات المختلفة بالمحافظات والبالغ عدد مفردات الدراسة من واقع سجلات شئون العاملين وهم (50) مفردة . وكذلك إدارة مبنى سنترال القبة والبالغ عدد مفردات الدراسة من واقع سجلات شئون العاملين وهم (360) مفردة . وبالتالي يكون اجمالى مفردات مجتمع الدراسة البالغ (500) مفردة مقسمة كما فى جدول رقم (1) .

جدول رقم (1)

توزيع مفردات المجتمع على الإدارات

النسبة من المجتمع (%)	العدد	البيان
72	360	إدارة مبنى سنترال القبة
18	90	الإدارة العامة للسلامة والصحة المهنية وجميع الاقسام التابعة لها فى الفروع
10	50	الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى الفروع
100	500	المجموع

ويتضح من الجدول رقم (1) توزيع مفردات مجتمع الدراسة على الإدارات .

عينة الدراسة :

وتم سحب عينة طبقية تمثل (55%) من مجتمع الدراسة اى تمثل (275) مفردة منهم (60) مفردة من الادارة العامة للسلامة وجميع الاقسام التابعة لها ومنهم ايضا (35) مفردة من الادارة العامة لإدارة وتحليل الازمات وايضا (180) مفردة من العاملين بإدارة مبنى سنترال القبة كما فى جدول رقم (2) .

وسوف يتم نشر (275) إستمارة إستقصاء صالحة للتحليل الاحصائى وذلك نظراً لضيق الوقت ولعدم إمكانية نشر (500) إستمارة وإعادة جمعهم وتحليلهم وإستخراج النتائج والتوصيات

جدول رقم (2)**توزيع مفردات العينة على الإدارات**

النسبة من مفردات المجتمع (%)	النسبة من العينة (%)	العدد	البيان
50	65	180	إدارة مبنى سنترال القبة
67	22	60	الادارة العامة للسلامة والصحة المهنية وجميع الاقسام التابعة لها فى الفروع
70	13	35	الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى الفروع
-----	100	275	المجموع

ويتضح من الجدول رقم (2) توزيع مفردات عينة الدراسة على الإدارات ونسبتها من مفردات المجتمع

متغيرات الدراسة

حيث تلاحظ ان متغيرات الدراسة متمثلة فى الاستبيان والذى يوضح مجموعتين من المتغيرات وهما متغيرات المجموعة الأولى (X) الخاصة بتطبيق قواعد واجراءات السلامة والصحة المهنية وتتكون من اربعة مؤشرات وهى كالتالى :

- التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية (X_1)
- الالتزام بتوفير وسائل السلامة والصحة المهنية في بيئة العمل (X_2)
- تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية (X_3)
- الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة والمهنية (X_4)

ومتغيرات المجموعة الثانية (Y) الخاصة بمراحل إدارة الازمات والكوارث وتتكون من خمسة مؤشرات وهى كالتالى:

- مرحلة اكتشاف إشارات الإنذار المبكر (y_1)
- مرحلة الاستعداد والوقاية (y_2)
- مرحلة احتواء الأضرار (y_3)
- مرحلة استعادة النشاط (y_4)
- مرحلة التعلم (y_5)

وفضلا عن ايضاحها في استمارة الاستبيان وكيفية قياس كل متغير فيها.

الاسلوب البحثي

أستخدم الأسلوب التحليلي في الدراسة حيث استخدم التحليل العاملي Factor Analysis لاستخلاص المتغيرات المستقلة والتابعة من عناصر صحيفة الاستبيان ثم تم حساب معاملات الارتباط البسيطة في شكل مصفوفة الارتباطات لتحديد شدة الارتباط بين المتغيرات المستقلة (قواعد واجراءات السلامة والصحة المهنية) والمتغيرات التابعة (مراحل إدارة الأزمة والكارثة). وحساب قيم Beta Coefficient التي تمثل معاملات الأنحدار في صيغتها القياسية.

نتائج الدراسة

هذا الجزء يستعرض اهم النتائج التي توصل اليها البحث وتتماشا مع الأهداف البحثية

جدول رقم (3)

الأوزان الناتجة من التحليل العاملي

المتغيرات (Variables)	الأوزان القياسية	العوامل (Factors)
• يوجد سياسة سلامة واضحة للعمل من قبل إدارة الشركة • يوجد تقييم وتحليل للمخاطر والكوارث الطبيعية والصناعية المتوقعة واعداد خطة طوارئ لحماية الشركة والعمال عند وقوع الكارثة من قبل الادارة • يوجد بالشركة اجهزة لقياس مخاطر بيئة العمل من قبل الادارة • هناك مشاركة للعاملين في عملية تقييم المخاطر وتطوير إجراءات السلامة والصحة المهنية • تشعر بوجود تعاون بينك وبين الإدارة لتحقيق تطبيق قواعد وإجراءات السلامة • يتم توفير الإمكانيات الفنية والمالية لتطبيق إجراءات السلامة والصحة المهنية • يوجد خطط وبرامج للسلامة المهنية والعمل على متابعتها	.654 .659 .229 .595 .673 .468	التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية ← (X_1)

	361. -115. 331. 553. 369. 729. 117.	• هناك تنسيق وتعاون من قبل الإدارة العليا مع المنظمات الأهلية بشأن تطبيق إجراءات السلامة والصحة المهنية • هناك وصف وظيفي لكل وظيفة يشمل تحديد إجراءات السلامة والصحة المهنية التي تنطوي عليها الوظيفة وإطلاع الموظف عليها. • يوجد كفاءات وخبرات في الإدارة العليا تستطيع أن تعمل على تطوير إدارة أنظمة السلامة والصحة المهنية • يوجد لجان للسلامة المهنية لتمارس دورها في تطوير إجراءات السلامة والصحة المهنية • تهتم الإدارة بالبحوث والدراسات العلمية التي تتعلق بتطوير إجراءات السلامة المهنية • هناك إدارة عامة بالشركة متخصصة في السلامة والصحة المهنية
الالتزام بتوفير وسائل السلامة والصحة المهنية في بيئة العمل ← (X_2)	375. 499. 818. 734. 259. -064. -110. -309. -029. 294. 229. -169. -280. -555. 126. 088.	• بناء السنترال (الجدران والأبواب والنوافذ) من مواد غير قابلة للاشتعال ومقاومة للمواد الكيميائية والإنفجارات • تجهيزات السنترال من (إضاءة – تهوية – حرارة – ضوضاء) تعتبر مناسبة • يهتم المتخصصون بالفحص الدوري (الماء ، الكهرباء ، الغاز) • من السهل الحصول على معدات السلامة الشخصية (معطف -قفاز -خوذة - نظارة -كمامة -أحذية - بطانية إطفاء -قناع الوجه) • هناك أنظمة إنذار متوفرة داخل السنترال و مناسبة لنوعية السنترال وبأعداد كافية وأماكن مناسبة. • يوجد ترخيص وسجل صيانة دورية للمصاعد • يوجد ملصقات ولوحات تحذيرية لتوعية العاملين بالسلامة في أماكن واضحة ومرئية • الطفايات اليدوية والمتحركة بأنواعها (لبودرة أو الرغوة أو السائلة أو الغازية)وإجمالي الرمل أعدادها كافية وأماكنها مناسبة. • كافة أجهزة وأدوات الإطفاء المستخدمة مطابقة للمواصفة القياسية المصرية • هناك صيانة دورية لطفايات الحريق وكذلك للأجهزة والمعدات من قبل مختصين • يوجد نظام يقوم بالفحص الطبي للعاملين الجدد • هناك سجل طبي للعاملين يتعلق بحوادث العمل داخل السنترال • يوجد مداخل ومخارج للطوارئ بالسنترال بأعداد كافية وأماكن مناسبة والممرات به ممهدة ومزودة بوسائل الإضاءة والاشارات • يوجد خزانات للإسعافات الأولية جاهزة للاستخدام في محيط العمل • وجود تعاون وتنسيق بين العاملين في السنترال في تطبيق وسائل السلامة • يوجد مشرف للسلامة والصحة المهنية والبيئة بكل سنترال
تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية ← (X_3)	584. 530. 506. 292. 581. 725. 421. 272.	• يتلقى العاملون تدريباً على خطة الإخلاء والطوارئ ومكافحة الحريق • يتم اختبار فاعلية خطط الطوارئ وإدارة الأزمات والكوارث الصناعية والطبيعية وأجراء تدريبات عملية عليها للتأكد من كفاءتها بصفة دورية • يتم التدريب على طرق التعامل مع أنظمة وأجهزة السلامة بالسنترال • هناك تدريب للعاملين على الإسعافات الأولية وكيفية استخدام مهمات الوقاية الشخصية أيضاً • يتم التدريب على إجراءات السلامة عن طريق (التدريب العملي-المحاضرات - النشرات المطبوعة- ورش العمل) • يتم توعية وتدريب العاملين بالمخاطر التي يواجهونها والزامهم باستخدام وسائل الوقاية المقررة • هناك مشاركة للعاملين في الندوات والمؤتمرات الدولية المتعلقة بالسلامة والصحة المهنية (منظمة العمل الدولية -منظمة الصحة العالمية) • تلقت تدريب على إجراءات السلامة من قبل جهات خارجية(الدفاع المدني - وزارة الصحة -وزارة العمل -مؤسسات أهلية)
الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة والمهنية ← (X_4)	541. 162. -454. -195. -037.	• هناك لوائح وقوانين خاصة بالسلامة والصحة المهنية داخل الشركة • توجد قوانين الزامية لاستخدام معدات الوقاية الشخصية • يتوفر منشورات وكتيبات إرشادية لكيفية استخدام الوسائل الوقائية • يتوفر داخل الشركة لوائح وقوانين تهتم بإجراءات السلامة والصحة المهنية • هناك التزام من قبل الشركة باللوائح والقوانين المصرية (قانون العمل المصري -قانون الصحة العامة) • هناك توجيهات وإرشادات للوائح المرتبطة بتطبيق الأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية

	• هناك التزام من قبل العاملين باللوائح والقوانين المتعلقة بقواعد السلامة المهنية • تفرض عقوبات رادعة على غير الملتزمين باستخدام وسائل الوقاية داخل السنترال • اللوائح والأنظمة الموجودة متوافقة مع المعايير الدولية (OSHA) و واضحة ومحددة • هناك قواعد وقوانين تطبق مبدأ التعويض والتأمين للعاملين	.857 .829 .300 .279 .153
مرحلة اكتشاف إشارات الإنذار المبكر ← (y_1)	• تولى الإدارة العليا في الشركة اهتماما ودعما لرصد مؤشرات حدوث الأزمات. • يتم مسح بيئة عمل الشركة بصورة شاملة للتعرف على مؤشرات احتمال حدوث الأزمة. • تهتم إدارة الشركة بعمليات تصنيف وتبويب مؤشرات حدوث الأزمات.	.798 .920 .816
مرحلة الاستعداد والوقاية ← (y_2)	• يتم تشكيل فرق مختلفة ومتعددة لحل العديد من الأزمات المحتملة في الشركة • يسهل الحصول على الإمكانات المادية والبشرية والتقنية والمعلوماتية المطلوبة من الإدارات والأقسام الأخرى عند الحاجة لها من أجل التعامل مع الأزمات. • تتوفر برامج وخطط كافية وجاهزة لإدارة الأزمات في الشركة ويتم العمل على مراجعتها وتطويرها باستمرار.	.552 .793 .839
مرحلة احتواء الأضرار ← (y_3)	• تقوم إدارة الشركة بالاستجابة الفورية لاحتواء الأزمة عن طريق توزيع المهام وتحديد الصلاحيات بفترة قصيرة ومناسبة عند حدوث الأزمة. • يتم التأثير في الأحداث واستخدام إجراءات الطوارئ التي تقلل وتحد من الأضرار التي تسببها الأزمة بكفاءة. • تقوم إدارة الشركة بإعداد غرفة عمليات مناسبة ومجهزة بالتقنيات الحديثة لاحتواء أسباب وأضرار الأزمة.	.650 .925 .858
مرحلة استعادة النشاط ← (y_4)	• تحافظ إدارة الشركة في ظروف الأزمات على اتخاذ الإجراءات اللازمة لمواصلة ممارسة النشاطات الاعتيادية في الشركة دون أي تأخير. • تقوم إدارة الشركة بتحديد الاحتياجات اللازمة للمواقع المختلفة التي تأثرت بالأزمة لمعالجة تأثيرات الأزمة واستعادة النشاط الاعتيادي. • تعمل إدارة الشركة على اتخاذ كافة الإجراءات اللازمة للتخفيف من آثار الأزمة والحد من استمرار حدوثها.	.563 .842 .729
مرحلة التعلم ← (y_5)	• تستخلص إدارة الشركة بصورة فعالة الدروس والعبر من الأزمات التي واجهتها سابقا في محاولة للاستفادة منها مستقبلا. • تقوم إدارة الشركة بتقييم خطط وبرامج إدارة الأزمات السابقة بقصد تطويرها وتحسينها من أجل التعامل مع الأزمات المستقبلية. • تعمل إدارة الشركة على الاستفادة من أساليب معالجة الأزمات في الإدارات الأخرى في البلد أو في الدول الأخرى ذات التجارب المشابهة.	.811 .848 .686

ويتضح من جدول رقم (3) الأوزان الناتجة من التحليل العاملي ،

و كانت اهم النتائج تمثلت في:

- أعلى الاوزان القياسية في مؤشر التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية (X_1) هو 729. والتي تمثل بعبارة تهتم الإدارة بالبحوث والدراسات العلمية التي تتعلق بتطوير إجراءات السلامة المهنية .
- أعلى الاوزان القياسية في مؤشر الالتزام بتوفير وسائل السلامة والصحة المهنية في بيئة العمل (X_2) هو 818 . والتي تمثل بعبارة يهتم المتخصصون بالفحص الدوري (الماء ، الكهرباء ، الغاز)

- أعلى الاوزان القياسية فى مؤشر تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية (X_3) هو 725. والتي تمثل بعبارة يتم توعية وتدريب العاملين بالمخاطر التي يواجهونها والزامهم باستخدام وسائل الوقاية المقررة
- أعلى الاوزان القياسية فى مؤشر الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية (X_4) هو 857. والتي تمثل بعبارة هناك توجيهات وإرشادات للوثائق المرتبطة بتطبيق الأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية
- أعلى الاوزان القياسية فى مؤشر مرحلة اكتشاف إشارات الإنذار المبكر (y_1) هو 920. والتي تمثل بعبارة يتم مسح بيئة عمل الشركة بصورة شاملة للتعرف على مؤشرات احتمال حدوث الأزمة.
- أعلى الاوزان القياسية فى مؤشر مرحلة الاستعداد والوقاية (y_2) هو 839. والتي تمثل بعبارة تتوفر برامج وخطط كافية وجاهزة لإدارة الأزمات في الشركة ويتم العمل على مراجعتها وتطويرها باستمرار .
- أعلى الاوزان القياسية فى مؤشر مرحلة احتواء الأضرار (y_3) هو 925. والتي تمثل بعبارة يتم التأثير في الأحداث واستخدام إجراءات الطوارئ التي تقلل وتحد من الأضرار التي تسببها الأزمة بكفاءة.
- أعلى الاوزان القياسية فى مؤشر مرحلة استعادة النشاط (y_4) هو 842. والتي تمثل بعبارة تقوم إدارة الشركة بتحديد الاحتياجات اللازمة للمواقع المختلفة التي تأثرت بالأزمة لمعالجة تأثيرات الأزمة واستعادة النشاط الاعتيادي .
- أعلى الاوزان القياسية فى مؤشر مرحلة التعلم (y_5) هو 848. والتي تمثل بعبارة تقوم إدارة الشركة بتقييم خطط وبرامج إدارة الأزمات السابقة بقصد تطويرها وتحسينها من أجل التعامل مع الأزمات المستقبلية.

جدول رقم (4) مصفوفة الارتباطات البسيطة بين متغيرات الدراسة										
		X_1	X_2	X_3	x_4	y_1	y_2	y_3	y_4	y_5
X_1	Pearson Correlation	1	-.083	.408	.384	.402	-.371	.276	.162	-.234
	Sig. (2-tailed)		.171	.000	.000	.000	.000	.000	.007	.000
X_2	Pearson Correlation	-.083	1	-.040	.037	.030	.052	-.205	.162	.181
	Sig. (2-tailed)	.171		.510	.537	.626	.393	.001	.007	.003
X_3	Pearson Correlation	.408	-.040	1	.098	-.089	-.208	.258	.161	-.286
	Sig. (2-tailed)	.000	.510		.103	.139	.001	.000	.007	.000
x_4	Pearson Correlation	.384	.037	.098	1	.328	-.237	.053	-.110	-.021
	Sig. (2-tailed)	.000	.537	.103		.000	.000	.381	.069	.723
y_1	Pearson Correlation	.402	.030	-.089	.328	1	-.099	-.142	.157	.061
	Sig. (2-tailed)	.000	.626	.139	.000		.103	.018	.009	.316
y_2	Pearson Correlation	-.371	.052	-.208	-.237	-.099	1	.137	.186	.260
	Sig. (2-tailed)	.000	.393	.001	.000	.103		.023	.002	.000
y_3	Pearson Correlation	.276	-.205	.258	.053	-.142	.137	1	-.098	-.450
	Sig. (2-tailed)	.000	.001	.000	.381	.018	.023		.106	.000
y_4	Pearson Correlation	.162	.162	.161	-.110	.157	.186	-.098	1	.172
	Sig. (2-tailed)	.007	.007	.007	.069	.009	.002	.106		.004
y_5	Pearson Correlation	-.234	.181	-.286	-.021	.061	.260	-.450	.172	1
	Sig. (2-tailed)	.000	.003	.000	.723	.316	.000	.000	.004	

ويتضح من جدول رقم (4) انه يوجد ارتباط بين أغلب بين متغيرات قواعد واجراءات السلامة والصحة المهنية التابعة لإدارة السلامة والصحة المهنية والمتمثلة في اربعة مؤشرات وهم (X_1 X_2 X_3 X_4) ومتغيرات مراحل إدارة الازمات والكوارث التابعة لإدارة الازمات والكوارث والمتمثلة في خمسة مؤشرات وهم (y_1 y_2 y_3 y_4 y_5) ،

و كانت اهم النتائج تمثلت في:

- مرحلة اكتشاف اشارات الإنذار المبكر ترتبط ارتباط معنوى مع (التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية و الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية) وذلك نظرا لأن Sig. اقل من 05 .

- مرحلة الاستعداد والوقاية ترتبط ارتباطاً معنوي مع (التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية و تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية و الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية) وذلك نظراً لأن Sig. اقل من 05 .
- مرحلة احتواء الأضرار ترتبط ارتباطاً معنوي مع (التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية و الالتزام بتوفير وسائل السلامة والصحة المهنية في بيئة العمل و تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية) وذلك نظراً لأن Sig. اقل من 05 .
- مرحلة استعادة النشاط ترتبط ارتباطاً معنوي مع (التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية و الالتزام بتوفير وسائل السلامة والصحة المهنية في بيئة العمل و تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية) وذلك نظراً لأن Sig. اقل من 05 .
- مرحلة التعلم ترتبط ارتباطاً معنوي مع (التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية و الالتزام بتوفير وسائل السلامة والصحة المهنية في بيئة العمل و تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية) وذلك نظراً لأن Sig. اقل من 05 .

جدول رقم (5)

معاملات الانحدار القياسية - وترتيب أهميتها

المتغيرات	الالتزام الإداري العليا لتطبيق إجراءات السلامة والصحة المهنية	الالتزام بتوفير وسائل السلامة والصحة المهنية في بيئة العمل	تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية	الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية
مرحلة اكتشاف اشارات الإنذار المبكر	(الاول) .46	(الرابع) .05	(الثاني) -.29	(الثالث) .18
مرحلة الاستعداد والوقاية	(الاول) -.29	(الرابع) .03	(الثالث) -.08	(الثاني) -.18
مرحلة احتواء الأضرار	(الاول) .21	(الثاني) -.18	(الثالث) .17	(الرابع) -.04
مرحلة استعادة النشاط	(الاول) .22	(الثالث) .19	(الرابع) .10	(الثاني) -.21
مرحلة التعلم	(الثالث) -.15	(الثاني) .16	(الاول) -.22	(الرابع) .05

ويتضح من جدول رقم (5) معاملات الانحدار القياسية وترتيب أهميتها وذلك بإستخدام قيم **Beta Coefficient** ،

و كانت اهم النتائج تمثلت فى :

● **مرحلة اكتشاف اشارات الإنذار المبكر** أن العوامل التى تؤثر فيها وبترتيبها حسب الدرجة الاعلى فى الاهمية وهى التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية ثم يليها تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية ثم يليهم الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية ثم يليهم فى المرحلة الاخيرة الالتزام بتوفير وسائل السلامة والصحة المهنية فى بيئة العمل وذلك بحساب قيم Beta.

● **مرحلة الاستعداد والوقاية** أن العوامل التى تؤثر فيها وبترتيبها حسب الدرجة الاعلى فى الاهمية وهى التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية ثم يليها الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية ثم يليهم تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية ثم يليهم فى المرحلة الاخيرة الالتزام بتوفير وسائل السلامة والصحة المهنية فى بيئة العمل وذلك بحساب قيم Beta.

● **مرحلة احتواء الأضرار** أن العوامل التى تؤثر فيها وبترتيبها حسب الدرجة الاعلى فى الاهمية وهى التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية ثم يليها الالتزام بتوفير وسائل السلامة والصحة المهنية فى بيئة العمل ثم يليهم تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية ثم يليهم فى المرحلة الاخيرة الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية وذلك بحساب قيم Beta.

● **مرحلة استعادة النشاط** أن العوامل التى تؤثر فيها وبترتيبها حسب الدرجة الاعلى فى الاهمية وهى التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية ثم يليها الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية ثم يليهم الالتزام بتوفير وسائل السلامة والصحة المهنية فى بيئة العمل ثم يليهم فى المرحلة الاخيرة تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية وذلك بحساب قيم Beta.

● **مرحلة التعلم** أن العوامل التى تؤثر فيها وبترتيبها حسب الدرجة الاعلى فى الاهمية وهى تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية ثم يليها الالتزام بتوفير وسائل السلامة والصحة المهنية فى بيئة العمل ثم يليهم التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية ثم يليهم فى المرحلة الاخيرة الالتزام بتوفير وتطبيق اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة المهنية وذلك بحساب قيم Beta.

خلاصة الدراسة

ويتضح ان التزام الإدارة العليا لتطبيق إجراءات السلامة والصحة المهنية لها الاولوية فى مراحل إدارة الازمات والكوارث ويليهها كلا من الالتزام بتوفير وسائل السلامة والصحة المهنية فى بيئة العمل و الالتزام بتوفير وتطبيق

اللوائح والأنظمة والقوانين المتعلقة بالسلامة والصحة والمهنية ولهما نفس الأهمية وثم يليهم في الأهمية المرحلة الأخيرة تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية

المراجع العربية

م	العربية
1	العباسي ، عبد الحميد ، (2016) " تحليل البيانات بإستخدام الحزم الإحصائية " معهد الدراسات والبحوث الإحصائية ، جامعة القاهرة .
2	الروسان وآخرون (2009) : الأمن الصناعي والسلامة المهنية ، الطبعة الثانية ، مكتبة المجتمع العربي للنشر والتوزيع ، عمان .
3	الشيخ خليل، علي (2008) : تقييم وسائل الوقاية والسلامة المستخدمة في مستشفيات قطاع غزة الحكومية وأثرها على أداء العاملين ، (رسالة ماجستير) ، الجامعة الإسلامية، غزة.
4	قطيشات، تالا وآخرون (2007) : مبادئ في الصحة والسلامة العامة ، الطبعة الثانية ، دار المسيرة للنشر والتوزيع ، عمان ، الأردن .
5	المغني، أميمه (2006) : واقع وإجراءات الأمن والسلامة المهنية المستخدمة في منشآت قطاع الصناعات التحويلية في قطاع غزة ، (رسالة ماجستير) ، الجامعة الإسلامية ، غزة.
6	المديفر، فهد (2005) : مدى فعالية تطبيق أنظمة الأمن والسلامة المهنية والتقنية ، دراسة مسحية على معامل الأقسام العلمية بكليات البنات ، الرياض (رسالة ماجستير) ، جامعة نايف العربية للعلوم الأمنية ، الرياض.
7	السرحي، حسن وآخرون (2002) : الصحة والسلامة في المكاتب المؤتمتة ، دراسة حالة على مكتبة جامعة الملك عبد العزيز ، مجلة الاتجاهات الحديثة في المكتبات والمعلومات بالقاهرة ، المجلد 7 ، العدد 13.
8	الشعلان ، فهد احمد (2002) . إدارة الأزمات : الاسس – المراحل – الآليات . (ط 2) ، الرياض " أكاديمية نايف العربية للعلوم الأمنية .
9	الهانسي ، مختار (1991) : " مقدمه في طرق التحليل الإحصاء " قسم الإحصاء والرياضة والتأمين ، كلية التجارة ، جامعة الاسكندرية
10	الحملوي، محمد رشاد (1995) إدارة الأزمات : تجارب محلية وعالمية" ، مكتبة عين شمس، القاهرة، ط ٢
11	الحملوي، محمد رشاد (1995) التخطيط لمواجهة الأزمات : عشر كوارث مرزت مصر" ، القاهرة، مكتبة عين شمس.
12	عليوة، السيد (2003) : إدارة الوقت والأزمات والإدارة بالأزمات ، القاهرة، دار الأمين للنشر والتوزيع.
13	برى ، عدنان و هندی ، محمود (2009) : " مبادئ الإحصاء والاحتمالات " ، مكتبة الشقري للنشر والتوزيع

14	طبية ، احمد عبدالسميع (2008) : " مبادئ الاحصاء " ، دار البداية ، عمان ، ط1
15	الهانسي ، مختار (1991) : " مقدمه فى طرق التحليل الاحصاء " قسم الاحصاء والرياضة والتأمين ، كلية التجارة ، جامعة الاسكندرية
16	طه ، محمد عبدالحميد (1985) " مقدمه فى الاحصاء " ، الهيئة العامة للتعليم التطبيقي والتدريب
17	محمد ، على ابو القاسم (1987) " اساليب الاحصاء التطبيقي " ، المعهد العربى للتخطيط بالكويت، دار الشباب للنشر والترجمة
18	منظمة العمل الدولية (2007)، أماكن عمل آمنة وصحية: جعل العمل اللائق واقعاً حقيقياً. مكتب العمل الدولي، جنيف .
19	عبد الرؤوف، مصطفى (2002) : الأمن والسلامة في المباني التعليمية واستخدام الحاسب الآلي في تقييم مستوى السلامة لمدارس التعليم الأساسي ، (رسالة دكتوراه) ، جامعة القاهرة ، مصر .
20	وزارة القوى العاملة (2003) : قانون العمل المصري رقم (12) لعام (2003) ،

المراجع الأجنبية

21	Bartlett, M.S. (1939), "A Note On Test Of Significance In Multivariate Analysis", Proceedings Of Cambridge Philosophical Society, 35.
22	Cristine Person and Ian Mitroff (1993): " From Crisis Prone To Crisis Prepared: A Framework For Crisis Management", Academy of Management Executive . Vol 7, No 1.
23	Graham Allison: Essence of Decision, U.S.A., Little Brown and Company, 1971, P.5.
24	Hotelling. H. (1936), "Relations Between Two Sets Of Variates", Biometrika, Vol28, Pp.321-336
25	Sterwart, D.K, And Love, W.A.(1968), "A General Canonical Correlation Index", Psychological Bulletin Vol,70, Pp.160-163.
26	Dejoy , D., Schaffer , B. & Wilson, N. , (2003) : " Creating Safer Workplaces : assessing the determinants and role of Safety Climate " , Journal of Safety Research , USA , 2003.
27	Lin, J. & Mill, A. (2001). Measuring the occupational health and safety performance of construction companies in Australia, Facilities, vol. 19, No. 3/4, pp. 131-138

المواقع الالكترونية

28	http://www.ilo.org/global
29	http://www.alolabor.org
30	http://www.education.gov.bh -
31	http://www.safety-eng.com -
32	http://www.salama-libya.org -

مؤشرات إحصائية للحوادث المرورية باستخدام تحليل السلاسل الزمنية بمنطقة عسير السعودية للفترة من (1417-1434) محمد محمد أحمد المزاح¹

الملخص

تعتبر الحوادث المرورية من أكبر الأخطار التي تواجه المجتمع وتستنزف موارده وطاقاته البشرية، حيث تناول هذا البحث تحليلاً إحصائياً زمنياً لأعداد الحوادث المرورية بمنطقة عسير للفترة 1417-1434هـ، مع التركيز على تعيين التوزيعات الاحتمالية المناسبة للبيانات، ومن ثم تقدير الطور (phase) العشوائي المولد للسلسلة الزمنية بهدف التعرف على الوضع الراهن لأعداد الحوادث، والتطورات المستقبلية لها، وإظهار السلوك المستقبلي لهذه الطوريات بناءً على النموذج المتنبأ به.

وقد أظهر شكل الانتشار للسلسلة الزمنية وجود حركة اهتزازية أقرب ما تكون للاهتزازات التوافقية، والتي تمثل المركبة البورية، وعدم وجود المركبة الموسمية، وأن النموذج المناسب لعرض بيانات السلسلة الزمنية هو النموذج التجميعي. وأن مركبة الاتجاه العام كانت بصورة نماذج غير خطية، أي أن النموذج المتنبأ به هو النموذج "اللا خطي". وتعد نتائج هذا البحث تقيماً لواقع فترة زمنية مهمة في تاريخ الحوادث، ونسبة الإصابة بالمنطقة. وأن أعداد الحوادث المرورية تسير بشكل متنامي، ومن المتوقع أن يزيد عدد الحوادث إلى حوالي (66173) حادثة في حلول عام 1443، ومتناقصة لنسبة الإصابة بالإضافة إلى وجود عوامل لها دور فاعل ومؤثر في الحوادث المرورية، وأن موضوع البحث بجانبه النظري والتطبيقي يفتح مجالاً واسعاً أمام الباحثين لاستخدام نماذج إحصائية تنبؤية أخرى.

الكلمات المفتاحية:

التذبذبات التوافقية، الحوادث المرورية، السلسلة الزمنية اللاخطية، نسبة الإصابة، النموذج التجميعي.

مقدمة

إن المتأمل لسير الحوادث المرورية بالسعودية، يلاحظ أنها صارت تمثل هاجساً وقلقاً وشعباً مخيفاً لكافة أفراد المجتمع، بل أضحت واحدة من أهم المشكلات التي تستنزف الموارد المادية والطاقات البشرية، وتستهدف أهم مقومات الحياة وهو العنصر البشري. وتنبع أهمية البحث من خطورة الدور الذي تلعبه الحوادث المرورية في الحياة العامة. وتشير الدراسات المرورية ومنها ما أوضحه رصد وحدة التقارير الاقتصادية بصحيفة الشرق الأوسط من تراجع الحوادث في المملكة بنسبة 9% خلال العام 2014م وانخفاض الوفيات إلى 20.5 حالة وفاة يومياً، وتصدرت منطقة الرياض من حيث الأكثر حوادث، ومنطقة مكة المكرمة من حيث الأكثر حالات الوفاة وتلتها منطقة الشرقية ثم منطقة عسير، كما تصدرت منطقة مكة المكرمة من حيث عدد الإصابات (الشرق الأوسط، 2015).

وأوضحت الدراسة التي أجراها عصام وآخرون (2013) أن 80% من الحوادث المرورية تقع داخل المدن، في حين 18% فقط تقع خارج المدن أو على الخطوط السريعة، وقدرت الدراسة الخسائر الاقتصادية كاملة لوفيات الحوادث المرورية بحوالي 79.92 مليار ريال، في حين يقدر إجمالي تكاليف علاج الإصابات البسيطة بحوالي 170.73 مليون ريال، والإصابات البليغة بـ 135 مليون ريال، وتكاليف الإضرار بالملكيات "السيارات" 6.94 مليار ريال، أما إجمالي التكاليف الشاملة للحوادث المرورية تقدر بحوالي 87.17 مليار ريال (jazirahonline.com/news).

يُعد تحليل السلاسل الزمنية من الأساليب العلمية والإحصائية لتفسير سلوك الظاهرة، حيث أشار الباحثان النعيمي والشاروط (2000) بدراستهما استخدام التدخل في السلاسل الزمنية لتقدير عدد المصابين بالأورام الخبيثة بتأثير عامل التدخل لتحديد أفضل نموذج للتنبؤ بعدد المصابين، ولاحظا أن أعداد المصابين بالأورام الخبيثة غير مستقرة في الوسط الحسابي والتباين، وأن هناك اتجاهًا عامًا وازدحاماً في السلسلة بعد عام 1973. أيضاً استخدم العلكة (2009) السلاسل الزمنية للتنبؤ بأعداد طلاب الماجستير ونسبة الخريجين منهم في الجامعات السورية، والذي توصل إلى أن أعداد طلاب الماجستير تسير بشكل متنامي (متزايد) في حين نسبة الخريجين منهم تتضاءل بشكل كبير. وقام الجبوري (2009) بدراسة السلاسل الزمنية ثنائية المتغيرات للتنبؤ بنسبة التضخم وعلاقته بسعر صرف الدولار الأمريكي مقابل الدينار العراقي، وتوصل إلى أن متجه التضخم النقدي يتبع نموذج الانحدار الذاتي ثنائي المتغيرات غير المستقر من الرتبة الثانية. كما قام محمد (2010) باقتراح نموذج إحصائي باستخدام تحليل السلاسل الزمنية لحوادث الحركة المرورية في ولاية الخرطوم، حيث قام بتقسيم بيانات السلاسل الزمنية إلى ثلاث سلاسل زمنية الأولى حوادث مرورية بسيطة، والثانية جسيمة، والثالثة حوادث أموات، وقد توصل إلى أن السلسلة الزمنية تسير بشكل متزايد وأن بعضها مستقر والبعض الآخر غير مستقر. أما طعمة (2012) فقد حدد في دراسته أن أفضل نموذج للتنبؤ بأعداد المصابين بالأورام الخبيثة والذي أظهر تزايد أعداد المصابين بالأورام بتأثير عامل

¹ قسم الرياضيات- كلية العلوم والآداب - جامعة الملك خالد - المملكة العربية السعودية.

قسم الرياضيات والحاسوب - كلية العلوم - جامعة إب - الجمهورية اليمنية.

التدخل، وقد لوحظ أن أعداد المصايين بالأورام الخبيثة غير مستقرة في الوسط الحسابي، وأن هناك اتجاهًا عامًا واضحًا في السلسلة بعد عام 2009 و أن النموذج الملائم للبيانات هو نموذج الانحدار من الرتبة الثانية. وسيقدم هذا البحث عروضاً وصفية مبسطة وبشكل أساسي على بحث سلوك السلسلة الزمنية الخاصة بالحوادث المرورية. وتعتبر نتائج هذا البحث أداة للتعرف على واقع فترة زمنية في تاريخ الحوادث المرورية، ويهدف هذا البحث إلى التنبؤ المستقبلي بأعداد الحوادث، والإصابات، والتطورات المستقبلية.

1. بعض الأسس النظرية لتحليل السلاسل الزمنية

السلسلة الزمنية هي مجموعة المشاهدات مرتبة وفق حدوثها، فهي بذلك عبارة عن سجل تاريخي يتم اعتماده لبناء التوقعات المستقبلية (فاندل، 1992)، وتتميز أي سلسلة زمنية بأن بياناتها مرتبة بالنسبة للزمن، وأن المشاهدات المتتالية عادة ما تكون غير مستقلة، مما يساعد على التوصل إلى تنبؤات موثوقة بها (Kirchgssner and Wolters, 2007) والغرض الأساسي من تحليل السلسلة هو الوصول لنموذج مناسب لتقدير أو قياس التغيرات، وبالتالي دراسة علاقاتها بالظروف المختلفة، من خلال التخلص من آثار العوامل المؤثرة في التغيرات وخاصة الاتجاه العام والتغيرات الدورية (Quenouille, 1957 and Murray, 2003).

نموذج السلسلة الزمنية:

يفترض نموذج السلسلة الزمنية أن قيم السلسلة دالة في مجموعة من العناصر المكونة لها مع الزمن: $Y = f(T_t, S_t, C_t, I_t)$ وفي هذا السياق

يرمز نموذجان لتجسيد العلاقة بين العناصر المتداخلة والتي يتفاعلها لتشكيل السلسلة الزمنية (Larsen and Marx, 1993)، ومنها:

1. النموذج التجميعي: يقوم على افتراض أن كل قيمة لكل من مكونات السلسلة لا تؤثر ولا تتأثر بقيمة غيرها من المكونات، ويكتب بالصيغة:

$$Y = T_t + S_t + C_t + I_t$$

2. النموذج الضربي: يقوم على افتراض أن قيم الظاهرة تساوي حاصل ضرب مكوناتها وأنها تعتمد على بعضها البعض أي:

$$Y = T_t \cdot S_t \cdot C_t \cdot I_t \quad (\text{Maestas and Robert, 2000})$$

قياس أثر الاتجاه العام T في حالة الاتجاهات غير الخطية Non-Linear Trends

يمكن تعيين عنصر الاتجاه العام بعدة طرق مختلفة منها طريقة المربعات الصغرى (البلداوي، 1997)، وسنركز على النماذج اللاخطية لتقدير معالم مركبة

الاتجاه العام (1) $Y = ax^2 + bx + c \rightarrow$ حيث a, b, c معاملات النموذج اللاخطي (Box and Jenkins, 1976). ويتم

تعيين قيم a, b, c من خلال العلاقات الآتية:

$$\left. \begin{aligned} \sum y &= nc + b \sum x + a \sum x^2 \\ \sum xy &= c \sum x + b \sum x^2 + a \sum x^3 \\ \sum x^2 y &= c \sum x^2 + b \sum x^3 + a \sum x^4 \end{aligned} \right\} \rightarrow (2)$$

أي أن:

$$\hat{a} = \frac{n \sum x^2 y - (\sum x^2)(\sum y)}{n \sum x^4 - (\sum x^2)^2} \rightarrow (3)$$

$$\hat{b} = \frac{\sum xy}{\sum x^2} \rightarrow (4)$$

$$\hat{c} = \frac{\sum y \sum x^4 - \sum x^2 y \sum x^2}{n \sum x^4 - (\sum x^2)^2} \rightarrow (5)$$

ويشترط لتحليل أية سلسلة زمنية أن تكون ساكنة أو تتسم بالسكون أي تحقق الشروط: ثبات المتوسط الحسابي والتباين أي يحقق

: $Var(Z_t) = E[(Z_t - \mu)^2] = \sigma^2$ (Brock and Davis, 1991)، وقد أشار (شعراوي، 2005) وأيضاً (Shumway and

Stoffer, 2006; Mosteller and Tukey, 1977; Clive and Granger, 2004) أن الفرضية الأساسية لتطبيق طريقة المربعات الصغرى في

النموذج الخطي هي عدم وجود ظاهرة الارتباط الذاتي. ونخلص مما سبق إلى أن الارتباط الذاتي يعتمد فقط على $(t-s)$ أي:

اختبار عشوائية أخطاء التنبؤ. $\rho_{t-s} = \frac{E[(Z_t - \mu)(Z_s - \mu)]}{\sigma^2} = \frac{\sigma_{ts}}{\sigma^2}$ وتستخدم دالة الارتباط الذاتي للبواقي (RACF) لفحص ملائمة النموذج عن طريق

نتائج تحليل السلاسل الزمنية

منهجية البحث

اعتمد البحث على بيانات السلسلة الزمنية لأعداد الحوادث المرورية والإصابات والتي تم الحصول عليها من الإدارة العامة للمرور بمنطقة عسير للفترة (1417-1434)، معتمداً المنهج الوصفي التحليلي من خلال الاطلاع على عدد من الأدبيات في هذا المجال، والتطبيق عليها وإجراء المعالجات الإحصائية

والعروض البيانية اللازمة من خلال بعض البرامج الحاسوبية والإحصائية (SPSS, Data Fit, Easy fit)

أولاً: أعداد الحوادث المرورية

بالرجوع إلى الإحصاءات السنوية الصادرة عن الإدارة العامة للمرور بمنطقة عسير خلال الفترة من (1417-1434) كانت بيانات الحوادث المرورية بمنطقة عسير كما هو موضح في جدول (1).

جدول (1): أعداد الحوادث المرورية خلال الفترة (1417-1434)

العام	1417	1418	1419	1420	1421	1422	1423	1424	1425
عدد الحوادث	7351	8095	7932	8274	9273	9521	10801	20042	18447
العام	1426	1427	1428	1429	1430	1431	1432	1433	1434
عدد الحوادث	22348	19550	20553	22954	25289	29469	27794	30530	31279

وبناءً على تحليل البيانات نلاحظ أن التوزيع الاحتمالي التوفيق Fitting of Distribution المناسب للبيانات المقدمة وفقاً لاختبار كيموجروف-سميرنوف هو توزيع Johnson SB Distribution

ذو المعالم $\lambda, \gamma, \delta, \zeta$ ، وأن دالة الكثافة الاحتمالية لهذا التوزيع تعطى بالصورة الآتية:

$$f(x) = \frac{1}{\lambda \sqrt{2\pi} z(1-z)} \exp \left[-\frac{1}{2} \left(\gamma + \delta \ln \left(\frac{z}{1-z} \right) \right)^2 \right] \rightarrow (6)$$

حيث $Z = \frac{X - \zeta}{\lambda}$ وأن مدى التوزيع الاحتمالي هو: $\zeta < X < \zeta + \lambda$ وأشارت البيانات تحت الدراسة أن:

$$\lambda = 25230, \zeta = 5955.4, \delta = 0.4322, \gamma = 0.02976,$$

ومن الجدول (2) نلاحظ أن توزيع بيانات الحوادث المرورية شبه ملتو ناحية اليمين التواء بسيطاً موجباً، وهذا يعني أن عدد الحوادث المسجلة في السنوات الأولى للفترة قيد البحث سار بشكل رتيب حول قيمة البدء، ومن ثم سار بشكل متذبذب في الفترات اللاحقة، ثم سار على نحو مطرد في السنوات الأخيرة للفترة الزمنية قيد البحث، أما من حيث التفرطح فإننا نلاحظ تدبياً قليلاً لتوزيع البيانات قيد الدراسة وهذا يعني أن تشتت البيانات حول وسطها الحسابي قليل نسبياً.

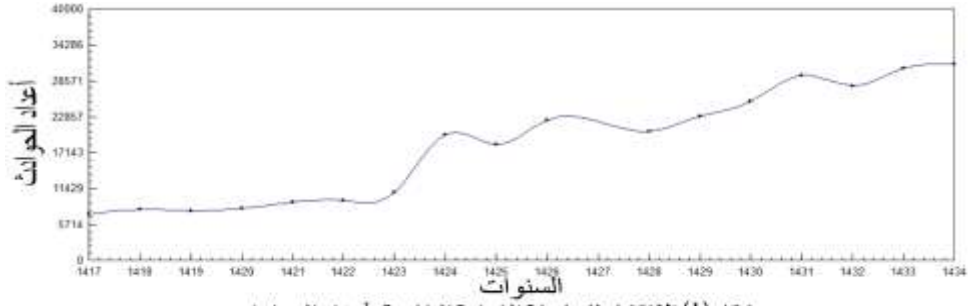
جدول (2): المقاييس الوصفية للبيانات الخاصة بالحوادث المرورية

اصغر قيمة	أكبر قيمة	المجموع	المتوسط الحسابي	الانحراف المعياري	معامل الالتواء	معامل التفرطح
7351	31279	329502	18305.67	8664.76	0.045	-1.528

التحليل الزمني للبيانات

إن دراسة السلسلة الزمنية تهدف إلى الحصول على وصف دقيق للملامح الخاصة بالطور (phase)

العشوائي المولد للسلسلة الزمنية بغرض بناء نموذج رياضي لتفسير وشرح سلوك السلسلة، بالاعتماد على معلومات الماضي، كما يفيد النموذج الرياضي الذي سيتم بناءه في تحديد القوى المؤثرة في السلسلة الزمنية بغرض الحصول على تنبؤات أكثر دقة، ومن المعلوم أن كل سلسلة زمنية سيرافقها طور عشوائي $\{\zeta_t\}_{t \in T}$ حيث يمكن تمثيل قيم هذه السلسلة الزمنية لقيم الطور العشوائي والتي يرمز لها بالرمز X_t من أجل $t \in T$ (العكلة، 2009)، وفي هذا البحث ستقوم بدراسة السلاسل الزمنية والتي أخذت مشاهداتها على فترات زمنية متساوية الطول (عام كامل) بحسب طبيعة البيانات قيد الدراسة، والعرض البياني الممثل بالشكل (1) يوضح انتشار السلسلة الزمنية الخاصة بأعداد الحوادث، حيث نلاحظ من الشكل وجود اتجاه عام Trend للسلسلة والذي يمثل تغيرات قيم الطور العشوائي المولد للسلسلة الزمنية على المدى البعيد.



شكل (1) الانتشار للسلسلة الزمنية الخاصة بأعداد الحوادث

وبعد أسلوب الانحدار من أكثر هذه الطرق شيوعاً في تحديد السلسلة الزمنية، واختبار الارتباط العشوائي للبيانات قيد الدراسة سوف نستخدم طريقة الفروق الأولى لـ موريس Wallis-Moore عند مستوى معنوية $\alpha = 0.05$ ، لإزالة أثر الاتجاه العام (Clive and Granger, 2004; George and Canavos, 1984).

فرضية العدم: H_0 : البيانات قيد البحث تمثل بيانات مستقلة عشوائياً أي أن: $\rho = 0$.

الفرضية البديلة: H_1 : البيانات قيد البحث بيانات مرتبطة عشوائياً أي أن: $\rho > 0$.

وللتحقق من ذلك تم إنشاء الجدول (3) الذي يعرض الفروق الأولى (الفروق المتتالية)، والتي تتم بطرح قيم مشاهدات السلسلة من بعضها البعض في ترتيب زمني محدد (العكس، 2009). والذي يبين أن عدد الإشارات الموجبة (+) يساوي 13 (كما علمنا بأن ζ هو متغير عشوائي يتوزع طبيعياً

$\zeta \sim N\left(\frac{n-1}{2}, \frac{n+1}{12}\right)$ (Larsen and Marx, 1993)، وحيث $(n < 30)$ فإن الإحصاء الخاص بهذا الاختبار نحصل عليه من:

$$TG = \frac{\left|\zeta - \frac{n-1}{2}\right| - \frac{1}{2}}{\sqrt{\frac{n+1}{12}}} = \frac{\left|13 - \frac{18-1}{2}\right| - \frac{1}{2}}{\sqrt{\frac{18+1}{12}}} = 3.1789 \rightarrow (7)$$

وبما أننا لاحظنا وجود اتجاه عام للسلسلة الزمنية فإننا سوف نستخدم اختبار ذو اتجاه واحد، وتكون المنطقة الحرجة عند مستوى معنوية 5% هي

1.64

وبالمقارنة نجد أن $TG < z_{0.05}$ وبناء على ذلك نرفض الفرضية الصفرية H_0 ونقبل بالفرضية البديلة H_1 عند مستوى معنوية $\alpha = 0.05$ أي أن البيانات قيد الدراسة مرتبطة عشوائياً.

جدول (3): الفروق الأولى (الفروق المتتالية)

العام	1417	1418	1419	1420	1421	1422	1423	1424	1425
عدد الحوادث المروية	7351	8095	7932	8274	9273	9521	10801	20042	18447
$Sig(x_{t+1}, x_t)^*$		+	-	+	+	+	+	+	-
العام	1426	1427	1428	1429	1430	1431	1432	1433	1434
عدد الحوادث المروية	22348	19550	20553	22954	25289	29469	27794	30530	31279
$Sig(x_{t+1}, x_t)^*$	+	-	+	+	+	+	-	+	+

وبعد تطبيق دالة الارتباط الذاتي ACF بواسطة برنامج SPSS على بيانات الجدول (1) حصلنا على معاملات الارتباط الذاتي الموضحة في الجدول (4) وكذلك من شكل (2) نلاحظ أن دالة الارتباط الذاتي تتناقص بشكل ملحوظ وهذا دليل على هدوء السلسلة، واختبار معنوية معاملات الارتباط الذاتي عند الفجوة الأولى $k=1$ بفترة ثقة 95% نلاحظ أن القيمة الجدولية 1.96 والمحسوبة 3.43

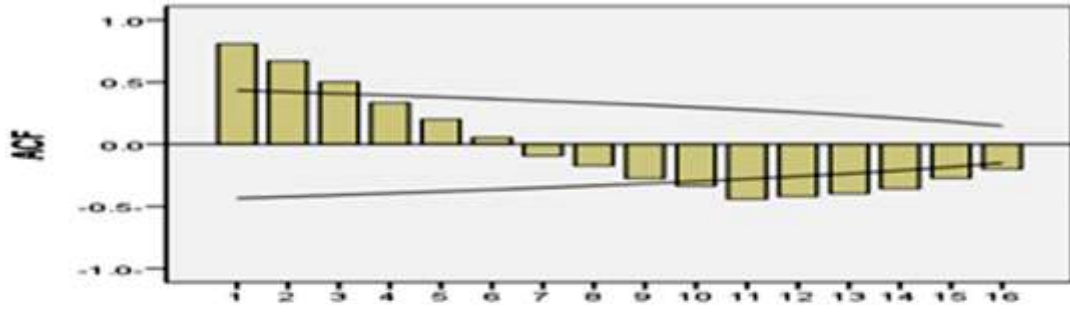
حيث $[Z_c = \frac{r}{\sqrt{1/n}}]$ وحيث أن القيمة المحسوبة أكبر من الجدولية نرفض فرضية انعدام معاملات الارتباط عند $k=1$. وعند اختبار

الفرضية عند $k=6$ نجد أن القيمة المحسوبة 0.229 وهذا يعني انعدام معاملات الارتباط الذاتي عند الفجوة 6، وبذلك يمكننا القول بهدوء

السلسلة الزمنية قيد البحث بعد الفجوة الخامسة. ومن شكل (1) نلاحظ ظهور حركة اهتزازية اقرب ما تكون للاهتزازات التوافقية والتي تمثلها المركبة الدورية للسلسلة الزمنية، في حين نرى أن المركبة الموسمية لا وجود لها في شكل الانتشار، لأنه من المعلوم بأن المركبة الموسمية تعبر عن التغيرات والتذبذبات الموسمية أو الفصلية الناتجة عن التغيرات في الفصول بسبب تأثير عوامل خارجية في مواسم محددة تعيد نفسها بشكل شبه منتظم على فترات زمنية متساوية الطول. وأن النموذج التجميعي هو النموذج المناسب عند انعدام إحدى المركبات (Granger,1977).

جدول (4) معاملات الارتباط الذاتي

Lag	1	2	3	4	5	6	7	8
Auto Corr	0.808 (0.000)	0.669 (0.000)	0.502 (0.000)	0.334 (0.000)	0.200 (0.000)	0.054 (0.000)	-0.092 (0.003)	-0.171 (0.000)
Lag	9	10	11	12	13	14	15	16
Auto Corr	-0.274 (0.000)	-0.335 (0.000)	-0.442 (0.000)	-0.419 (0.000)	-0.397 (0.00)	-0.357 (0.000)	-0.273 (0.000)	-0.200 (0.000)



شكل (٢) : دالة الارتباط الذاتي للبواقي للنموذج اللاخطي

تقدير عنصر الاتجاه العام للسلسلة الزمنية

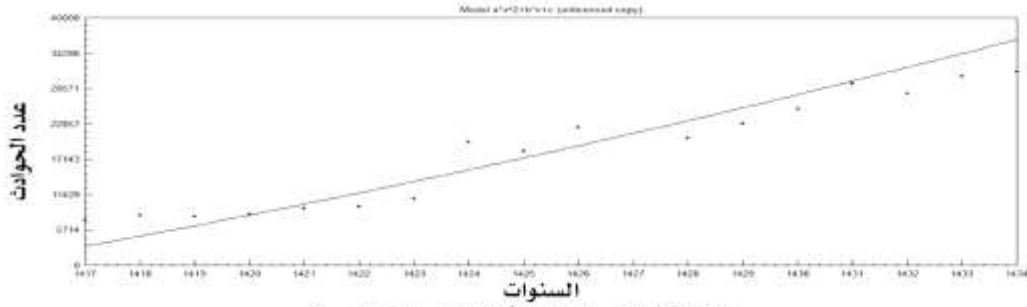
لتقدير عنصر الاتجاه العام $T(t)$ سوف نفترض ضمناً أن المكونين الدوري والموسمي محتواة في عنصر الخطأ (سواء كانت موجودة أم لا) وتكتب بالصورة الآتية: $Y_t = T(t) + I_t$ وفي هذه الحالة يجب أن نُحدد الخطأ العشوائي I_t أثناء عملية التقدير وبضاد كمتغير عشوائي بعد الانتهاء من عملية التقدير. ولذلك لن يتم إدخاله في عملية تقدير عنصر الاتجاه العام والعنصر الدوري، ويعتمد تقدير عنصر الاتجاه العام على التعرف على خط مستقيم يصف لنا حركة السلسلة الزمنية في فترة زمنية طويلة نسبياً، وبما أن البيانات قيد الدراسة هي بيانات سنوية فإن انحرافات قيم السلسلة عن خط الاتجاه العام لا تمثل فقط التغيرات العشوائية أو العرضية، بل تمثل التغيرات الدورية، وهذا يعني تعيين خط انحدار مناسب لبيانات السلسلة قيد الدراسة، وهذا يتطلب منا توفيق منحنى fitting curve لبيانات السلسلة الزمنية (Hamilton and James,1994). ولتوفيق هذا المنحنى نستخدم برنامج Curve(Data fit)، الذي يقوم بتوفيق المنحنى اعتماداً على طريقة المربعات الصغرى من جهة، ومن جهة أخرى يقوم البرنامج بتعيين المنحنى ذو الخطأ المعياري الأصغر أولاً، ومن ثم تعيين أكبر قيمة لمعامل الارتباط، حيث تم تقدير معالم نموذج الاتجاه التربيعي، فقد أوضحت نتائج التحليل أن القيم التقديرية لمعاملات النموذج في (3,4,5) a, b, c هي:

$$\hat{a} = 13.416, \hat{b} = 3667.719, \hat{c} = 25038286.026$$

أي يأخذ النموذج اللاخطي الصورة:

$$\hat{T}(t) = 13.416t^2 - 3667.719t + 25038286.026$$

ونحطاً معياري $S = 94.04$ ومعامل ارتباط $r = 0.967$ ومعامل التحديد $R^2 = 0.94$ ، وأن كفاءة النموذج حسب اختبار درين-وتسن الإحصائي هو: $d = 1.592$ وأن $d_u = 1.39$ ، $d_L = 1.16$ ، وحيث أن $d_U < d$ فلا يوجد لدينا ما يدفعنا لرفض الفرضية الصفرية H_0 عند مستوى معنوية $\alpha = 0.05$ ، والشكل (3) يوضح مركبة الاتجاه العام لهذه السلسلة الزمنية.



شكل (3) الاتجاه العام لأعداد الحوادث المرورية

تقدير العنصر الدوري للسلسلة الزمنية

يمكن تقدير العنصر الدوري بطريقة البواقي (Residual method) وذلك بفصله من عنصر الاتجاه العام وبالتالي يبقى اثر التغيرات الدورية لهذه

$$\zeta_t = \hat{T}(t) + C(t) + I(t) \rightarrow (8)$$

السلسلة، ويصبح الطور العشوائي المولد للسلسلة على الصورة:

$$\tilde{\zeta}_t = \zeta_t - \hat{T}(t) = C(t) + I(t) \rightarrow (9)$$

والذي يمكن كتابته بالصورة:

تسمى هذه العلاقة باسم النموذج المختزل للسلسلة الزمنية، وبموجب الفرضيات الخاصة بالعنصر العشوائي، وحيث ($n < 30$) واهتزازات هذه السلسلة أقرب ما تكون إلى الاهتزازات التوافقية وبأدوار متساوية الطول ودورة كل منها تساوي $P = 5$ على وجه التقريب (وهو عدد القياسات في كل دورة) وبناء على ذلك فإن الطريقة المناسبة لتقدير العنصر الدوري هي طريقة جيب التمام لـ هالبرغ (Halberg Cosine method) والتي تقوم على عدد من الفرضيات (Hamilton and James, 1994)، وتكتب بالصيغة الآتية

$$C(t) = \mu + A \cos(wt + \psi) \rightarrow (10)$$

حيث μ : هي القيمة المتوسطة لقيم الفروق Residual Values وأن قيمتها نظرياً $\mu = 0$ لأن المعاملات $\hat{a}$, $\hat{b}$, $\hat{c}$ اختيرت بشكل يحقق المعادلة (2) حتى وإن كانت قيمتها العددية $\mu \neq 0$ و A : سعة الاهتزاز و w : السرعة الزاوية للاهتزاز والتي تساوي $w = \frac{2\pi}{P}$ أي أن: $w = \frac{2\pi}{5} = 1.257$ و ψ : فرق المسار (الطور) بين السلسلة الزمنية ودالة جيب التمام، وحيث أن السرعة الزاوية للاهتزاز w معلوم دائماً من بيانات العينة المأخوذة من مسار السلسلة الزمنية. ويتم تقدير المعالم μ , A , ψ بالاعتماد على قيم الفروق $\tilde{X}_t$ بين القيم الفعلية X_t والقيم المقدرة $\hat{X}_t$ وفقاً للمنحنى التوافقي الخاص بالبيانات تحت الدراسة، وسنقدم هنا تقديرات لمعالم العنصر الدوري مباشرة دون اللجوء إلى العمليات الحسابية المطولة، وبفرض n حجم العينة المأخوذة من مسار السلسلة الزمنية وأن القيمة المقدرة لـ μ , α , ψ هي:

$$\mu = -1.36 \times 10^{-7}, A = 93.524, \psi = -1.431$$

علماً بأن $\mu = -1.36 \times 10^{-7}$ تعتبر تافهة بالمقارنة مع بقية القيم

وبالتعويض نجد أن الدالة المقدرة للمركبة الدورية $\hat{C}(t)$ هي:

$$\hat{C}(t) = -1.36 \times 10^{-7} + 93.524 \cos(\pi/2t - 1.431)$$

وبناءً على ما سبق نلاحظ أن الصيغة المثالة للنموذج المقدّر للطور (phase) العشوائي المولد للسلسلة الزمنية لأعداد الحوادث المرورية يأخذ الصيغة الآتية:

$$\hat{\zeta}_t = 13.416t^2 - 36676.862t + 25038387.436 - 1.36 \times 10^{-7} + 93.524 \cos(\pi/2t - 1.431) + I(t)$$

ومن خلال استعراضنا لأحد المسارات المقدّرة للطور العشوائي المولد للسلسلة المعطاة نلاحظ أن القيمة المقدّرة للعام 1425 و 1434 وفقاً للنموذج العشوائي المقدّر للطور العشوائي المولد للسلسلة الزمنية هي على الترتيب (17281 و 32050) حيث كان الخطأ المعياري $S = 101.83$ والقيم الحقيقية لهذين العامين هي (19342 و 31279) ما زالتا تقعان ضمن التقدير الجيد، وفقاً للطور العشوائي المقدّر الذي قمنا بتعيينه، مما يزيد ثقتنا في استخدامنا لهذا النموذج في التنبؤ، والشكل (4) يوضح ذلك:



يمكن استخدام النموذج المقدر للطور العشوائى المولد للسلسلة للتنبؤ بمشاهدات السلسلة مستقلاً، ويعتبر التنبؤ في هذه الحالة من المستوى طويل الأمد، والذي يغطي فترة زمنية أكثر من خمس سنوات، ولهذا فإنه وعلى المدى المسموح به للتنبؤ وفي هذه الحالة (يجوز لنا استخدام هذا المنحنى للتنبؤ حتى نهاية فترة طولها يساوي نصف طول الفترة التي استخدمت للدراسة) ومن الشكل (5) يلاحظ أن عدد الحوادث المرورية في المنطقة تظهر تنامياً (تزايداً) واضحاً حتى على مدى السنوات المقبلة، والذي من المتوقع أن يزيد العدد إلى حوالي (49090) حادثة في عام 1443هـ.



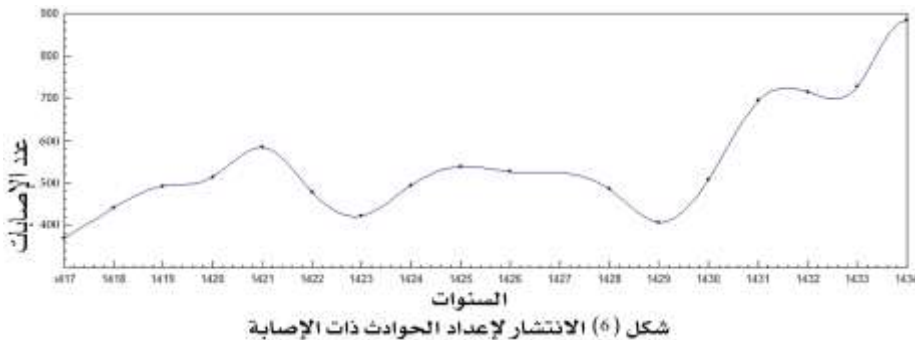
ثانياً: التحليل الزمني لأعداد الحوادث ذات الإصابات

بالرجوع إلى الإحصاءات السنوية الصادرة عن الإدارة العامة للمرور بمنطقة عسير خلال الفترة من (1417-1434) فإن الجدول (5) يبين الإحصاءات السنوية لأعداد الحوادث المرورية ذات الإصابات لمنطقة عسير

جدول (5): أعداد الحوادث المرورية ذات الإصابات

العام	1417	1418	1419	1420	1421	1422	1423	1424	1425
عدد الإصابات	369	441	492	513	584	479	422	494	538
العام	1426	1427	1428	1429	1430	1431	1432	1433	1434
عدد الإصابات	526	470	485	407	508	694	714	727	884

العرض البياني الممثل بالشكل (6) يوضح الانتشار للسلسلة الزمنية الخاصة بأعداد الحوادث المرورية التي حدث فيها إصابات، يشير إلى وجود اتجاه عام Trend متزايد للسلسلة الزمنية يتوافق مع السلوك العام لأعداد الحوادث المرورية. حيث نلاحظ ظهور حركة اهتزازية اقرب ما تكون للاهتزازات التوافقية والتي تمثلها المركبة الدورية للسلسلة الزمنية، في حين أن المركبة الموسمية لا وجود لها في شكل الانتشار، وأن النموذج التجميعي هو النموذج المناسب لعرض هذه السلسلة الزمنية (Granger, 1977).



وبلاحظ من خلال الشكل (6) أن أعداد الإصابات بدءاً من العام 1430 بدأت تتضاعف عن الأعوام السابقة وهذا يعود إلى أن إدارة المرور قد طورت نظام رصد الحوادث من خلال أمن الطرقات وكمرات المراقبة، والدوريات الأمنية. واختبار الارتباط العشوائي للبيانات قيد الدراسة نستخدم طريقة الفروق الأولى عند مستوى معنوية $\alpha = 0.05$ لإزالة أثر الاتجاه العام. ومن جدول (6) نلاحظ أن عدد الإشارات الموجبة (+) يساوي $12 = K$ فإن قيمة الفروق المحسوبة لهذا الاختبار يمكن الحصول عليها من العلاقة رقم (7). وبما أننا لاحظنا وجود اتجاه عام للسلسلة الزمنية فإننا سوف نستخدم اختبار ذو اتجاه واحد، وتكون المنطقة الحرجة عند مستوى معنوية 5% هي 1.64

وبالمقارنة نجد أن $Z_\alpha < TG$ وبناء على ذلك نرفض الفرضية الصفريّة H_0 وقبل بالفرضية البديلة H_1 عند مستوى معنوية $\alpha = 0.05$ أي أن البيانات قيد الدراسة مرتبطة عشوائياً، ويمكننا بحث هدوء (سكون) السلسلة الزمنية قيد الدراسة. جدول (6): الفروق الأولى (الفروق المتتالية)

العام	1417	1418	1419	1420	1421	1422	1423	1424	1425
عدد الإصابات	369	441	492	513	584	479	422	494	538
$sig(x_{t+1}, x_t)$		+	+	+	+	-	-	+	+
العام	1426	1427	1428	1429	1430	1431	1432	1433	1434
عدد الإصابات	526	470	485	407	508	694	714	727	884
$sig(x_{t+1}, x_t)$	-	-	+	-	+	+	+	+	+

ومن خلال تفحص معاملات الارتباط الذاتي $\hat{\rho}_k$ الموضحة في جدول (7) لبيانات العينة الممثلة لمسار السلسلة الزمنية قيد الدراسة، وتطبيق دالة الارتباط الذاتي ACF بواسطة برنامج SPSS على بيانات الجدول (5) نلاحظ أن دالة الارتباط الذاتي تتناقص بشكل ملحوظ؛ وهذا دليل على هدوء السلسلة الزمنية، وإذا أردنا اختبار انعدام معاملات الارتباط الذاتي عند الفجوة الأولى $k = 1$ بفترة ثقة 95% تكون القيمة المحسوبة 3.084 والقيمة الجدولية 1.96، وحيث أن القيمة المحسوبة أكبر من القيمة الجدولية نرفض فرضية انعدام معاملات الارتباط عند $k = 1$ ، وبالمثل عند $k = 5$ نجد أن القيمة المحسوبة 0.127 وهذا يعني انعدام معاملات الارتباط الذاتي عند 5 درجات، وبذلك يمكننا القول بهدوء السلسلة الزمنية قيد البحث بعد الفجوة الرابعة، أي أن U_t مستقلة عن U_{t-1} .

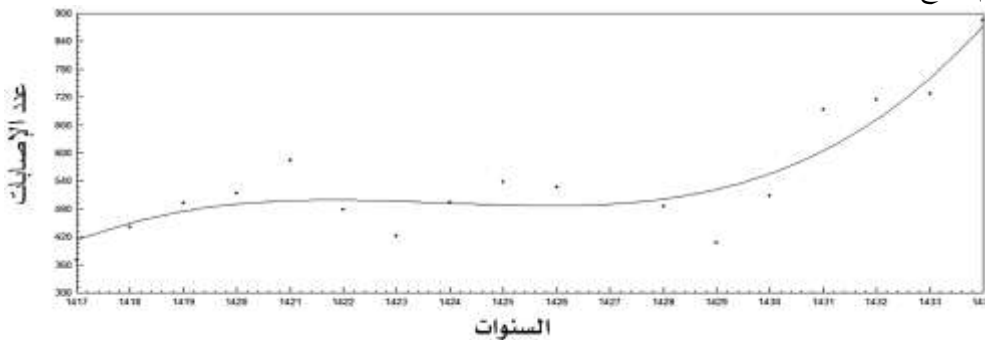
جدول (7): معاملات الارتباط الذاتي

Lag	1	2	3	4	5	6	7	8
Auto Corr	0.727 (0.000)	0.465 (0.000)	0.245 (0.001)	0.072 (0.001)	0.030 (0.003)	-0.015 (0.007)	-0.067 (0.013)	-0.094 (0.020)
Lag	9	10	11	12	13	14	15	16
Auto Corr	-0.110 (0.029)	-0.130 (0.036)	-0.164 (0.036)	-0.190 (0.028)	-0.214 (0.01)	-0.300 (0.002)	-0.299 (0.000)	-0.268 (0.000)

ولتقدير الطور (phase) العشوائي المولد للسلسلة الزمنية نقوم بتقدير العناصر التي يلاحظ وجودها.

تقدير عنصر الاتجاه العام للسلسلة الزمنية

بنفس الأسلوب السابق سنقوم بتقدير عنصر الاتجاه العام $T(t)$ ومن خلال الشكل (7) نلاحظ أن أحسن خط انحدار لبيانات السلسلة الزمنية باستخدام برنامج Curve(Data fit) هو:



شكل (7) مركبة الاتجاه العام لأعداد الحوادث ذات الاصابة

$$T(t) = at^4 + bt^3 + ct^2 + dt + e \rightarrow (11) \quad \text{والذي معادلته}$$

حيث أوضحت نتائج التحليل أن القيم التقديرية لـ a, b, c, d, e هي:

$$\hat{a} = 3.8441 \times 10^{-3}, \quad \hat{b} = -21.52086, \quad \hat{c} = 45167.999,$$

$$\hat{d} = -42119619.28, \quad \hat{e} = 1472412340.03$$

وبناء على ذلك فإن نموذج الاخطي يأخذ الصورة:

$$\hat{T}(t) = 3.8441 \times 10^{-3} t^4 - 21.52086 t^3 + 45167.999 t^2 - 42119619.28 t + 1472412340.03$$

ونخطأ معياري لمعادلة الانحدار مقداره: $S = 83.04$ ومعامل ارتباط $r = 0.918$ ومعامل التحديد $R^2 = 0.842$ ، وأن كفاءة النموذج حسب اختبار ديرين-وتسن الإحصائي هو: $d = 1.56$ وأن $d_u = 1.39$ ، $d_L = 1.16$ ، وحيث أن $d_U < d$ ومن ثم فلا يوجد لدينا ما يدفعنا لرفض الفرضية الصفرية H_0 عند مستوى معنوية $\alpha = 0.05$. ومن العلاقة السابقة نلاحظ أن الاتجاه العام لهذه السلسلة الزمنية لها شكل دالة اهتزازية، ومتزايدة بدرجة ملحوظة خلال السنوات العشر القادمة.

تقدير العنصر الدوري للسلسلة الزمنية

وبشكل مماثل لما سبق يمكن تقدير العنصر الدوري، وتحت الفرضيات الخاصة بالعنصر العشوائي أن لهذا العنصر العرض الآتي وفقاً لطريقة التجيب لـ هالبرغ، وبافتراض أن العنصر الدوري العرض

$$p(t) = \mu + \alpha \cos(wt + \psi) \quad \text{ف نجد أن قيمة مقدر } \mu \text{ هي: } \hat{\mu} = -0.0001847 \text{ علماً بأن}$$

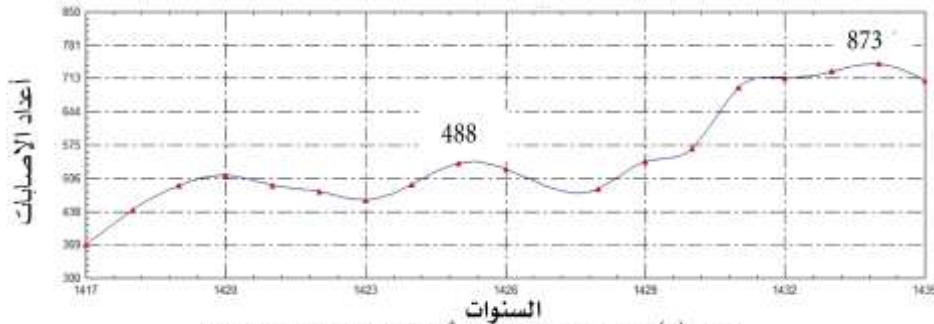
$$w = \frac{2\pi}{p} = 1.257 \quad \text{وأن قيمة } \hat{\alpha} = 18.735, \quad \hat{\psi} = 4.632 \quad \text{وبالتعويض في صيغة المركبة الدورية نجدها تعرض حسب الصيغة الآتية:}$$

$$\hat{p}(t) = -0.0001847 + 18.735 \cos(1.257t + 4.632)$$

وبالتعويض في المعادلة (8) نجد أن الصيغة المثلثة للنموذج المقدّر للطور (phase) العشوائي المولد للسلسلة الزمنية الخاص بأعداد الحوادث ذات الإصابات يأخذ الصورة الآتية:

$$\hat{G}_t = 3.8441 \times 10^{-3} T^4 - 21.52086 T^3 + 45167.999 T^2 - 42119619.28 T + 1472412340.3 - 0.0001647 + 18.735 \cos(1.257t + 4.632)$$

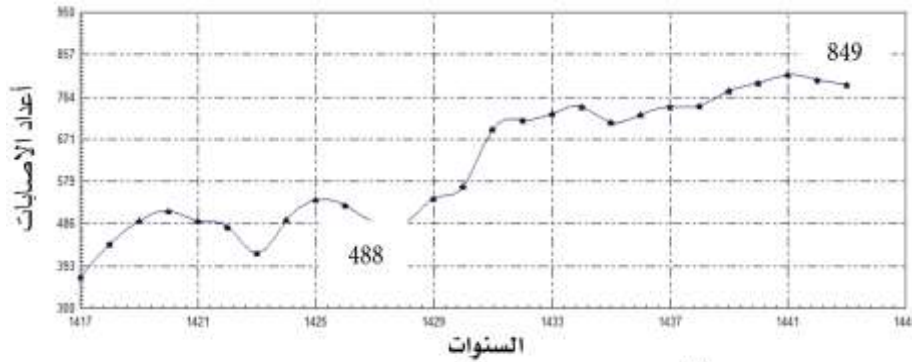
وبناء على ما سبق يصبح لأحد المسارات المقدرة للطور العشوائي المولد للسلسلة قيد البحث بالشكل (8)



شكل (8) أحد المسارات المقدرة لأعداد الحوادث ذات الإصابات

ومن خلال استعراضنا لأحد المسارات المقدرة للطور العشوائي المولد للسلسلة الزمنية الموضح في الشكل (9) نلاحظ أن القيم المقدرة ما زالت تقع ضمن التقدير الجيد، وفقاً للطور العشوائي المقدّر الذي قمنا

بتعيينه، مما يزيد ثقتنا في استخدامنا لهذا النموذج في التنبؤ.



شكل (9) النموذج المقدر للتنبؤ بأعداد الحوادث ذات الإصابات

ونلاحظ مما سبق أن أعداد المصابين في الحوادث المرورية بالمنطقة يظهر تنامياً (تزايداً) واضحاً حتى على مدى السنوات المقبلة، والذي من المتوقع أن يزيد العدد إلى حوالي (849) حادثة في حلول عام 1443هـ (بفرض استقرار الظروف).

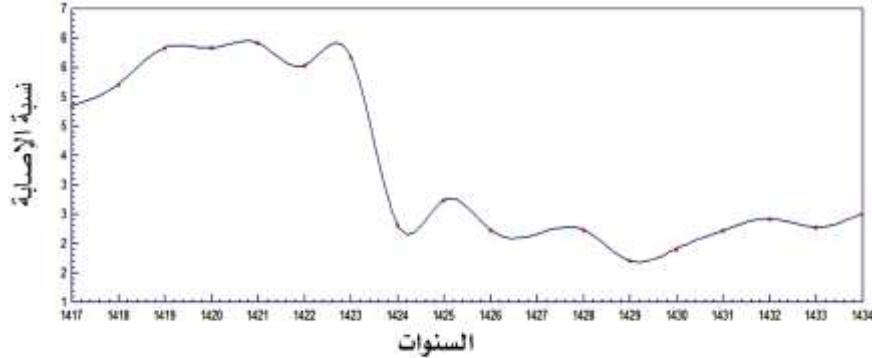
ثالثاً: التحليل الزمني للنسبة المئوية لأعداد الحوادث ذات الإصابات

من العوامل المهمة في مثل هذه المسائل معرفة النسب المئوية لأعداد الحوادث ذات الإصابات لأنها تعطينا تصوراً واضحاً حول طبيعة الحوادث في هذه المرحلة ومدى توفر وسائل الحماية والأمن والسلامة، ونوعية المركبة، وخدمات الطرق. والجدول (8) يشير إلى النسب المئوية لأعداد الحوادث ذات الإصابات.

جدول (8): النسب المئوية لأعداد الحوادث التي حدثت ذات الإصابات.

العام	1417	1418	1419	1420	1421	1422	1423	1424	1425
نسبة الإصابات	5.019	5.448	6.203	6.200	6.298	5.833	6.015	2.581	3.089
العام	1426	1427	1428	1429	1430	1431	1432	1433	1434
نسبة الإصابات	2.469	2.528	2.478	1.843	2.092	2.462	2.707	2.528	2.826

العرض البياني في الشكل (10) الخاص بنسبة الإصابة يشير إلى أن الاتجاه العام لهذه السلسلة الزمنية متذبذباً بمدى ثابت وتكرر ثابت، كما يبين ظهور حركة اهتزازية أقرب ما تكون للاهتزازات التوافقية. أي أنه يشير إلى أن النسبة المئوية لأعداد الحوادث ذات الإصابات بدأت من العام 1434هـ تتناقص عن الأعوام السابقة بشكل ملحوظ، وهذا قد يرجع إلى توافر بعض العوامل المؤثرة مثل وسائل الحماية وخدمات الطرق.



شكل (10) الانتشار لنسبة الحوادث ذات الإصابات

و يشير جدول (9) إلى الارتباط العشوائي للبيانات باستخدام طريقة الفروق الأولى لإزالة أثر الاتجاه العام.

جدول (9): الفروق الأولى (الفروق المتتالية)

العام	1417	1418	1419	1420	1421	1422	1423	1424	1425
نسب الإصابات	5.019	5.448	6.203	6.200	6.298	5.833	6.015	2.581	3.089
$sig(x_{t+1}, x_t)^*$		+	+	-	+	-	+	-	+
العام	1426	1427	1428	1429	1430	1431	1432	1433	1434
نسب الإصابات	2.469	2.528	2.478	1.843	2.092	2.462	2.707	2.528	2.826
$sig(x_{t+1}, x_t)^*$	-	+	-	-	+	+	+	-	+

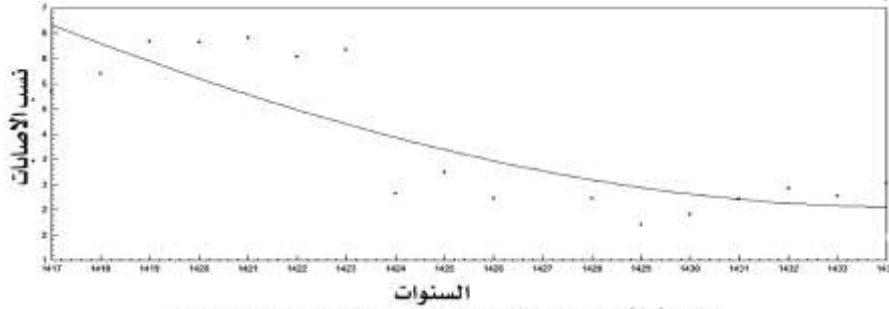
ومن الجدول (9) نلاحظ أن عدد الإشارات الموجبة (+) يساوي $10 = \zeta$ ، علماً بأن القيمة المحسوبة للفروق الأولى وفقاً للمعادلة (7) هي $TG = 0.7947$ وتكون المنطقة الحرجة عند مستوى معنوية 5% هي 1.64 . وبالمقارنة نجد أن $TG < Z_{\alpha}$ ومن ثم ليس لدينا ما يدفعنا لرفض الفرضية الصفرية H_0 عند مستوى $\alpha = 0.05$. وكما سبق ومن خلال تفحص معاملات الارتباط الذاتي المتناقصة بشكل ملحوظ كما في جدول (10)، وبذا نستطيع القول بهدوء السلسلة الزمنية.

جدول (10) : يوضح معاملات الارتباط الذاتي

Lag	1	2	3	4	5	6	7	8
Auto Corr	0.835 (0.000)	0.705 (0.000)	0.485 (0.001)	0.280 (0.000)	0.061 (0.000)	-0.138 (0.000)	-0.289 (0.000)	-0.338 (0.020)
Lag	9	10	11	12	13	14	15	16
Auto Corr	-0.372 (0.000)	-0.392 (0.000)	-0.390 (0.000)	-0.322 (0.000)	-0.246 (0.00)	-0.176 (0.000)	-0.115 (0.000)	-0.063 (0.000)

تقدير عنصر الاتجاه العام للسلسلة الزمنية

وعلى نحو مماثل لما سبق سنقوم بتقدير عنصر الاتجاه العام $T(t)$ فنلاحظ من خلال الشكل (11) أن أفضل خط انحدار لبيانات السلسلة الزمنية باستخدام برنامج Curve (Data fit) هو:



شكل (11) مركبة الاتجاه العام لنسبة الحوادث ذات الإصابة

والذي معادلته $T(t) = aX^3 + bX^2 + cx + d \rightarrow (12)$ حيث أوضحت نتائج التحليل أن القيم التقديرية هي: $\hat{d} = -436548.596$ ، $\hat{c} = 942.0416$ ، $\hat{b} = -0.67406$ ، $\hat{a} = 1.6066$ وبناء على ذلك فإن النموذج اللاخطي يأخذ الصورة: $\hat{T}(t) = 1.6066T^3 - 0.6741T^2 + 942.042T - 438548.596$ وبمعامل ارتباط $r = 0.837$ ومعامل التحديد $R^2 = 0.6998$ ، وأن كفاءة النموذج حسب اختبار ديرين-وتسون الإحصائي هو $d = 1.001$ وأن $d_u = 1.39$ ، $d_L = 1.16$ ، وحيث أن $d < d_L$ ومن ثم فلا يوجد لدينا ما يدفعنا لقبول الفرضية الصفرية H_0 عند مستوى معنوية $\alpha = 0.05$. أي أنه لو استمر عدد الإصابات بالحوادث المرورية على النحو الذي تم في الماضي فإن نسبة الإصابات ستتضاءل خلال العقد المقبل، نتيجة توافر بعض العوامل المؤثرة مثل وسائل الحماية، وخدمات الطرق.

تقدير العنصر الدوري للسلسلة الزمنية

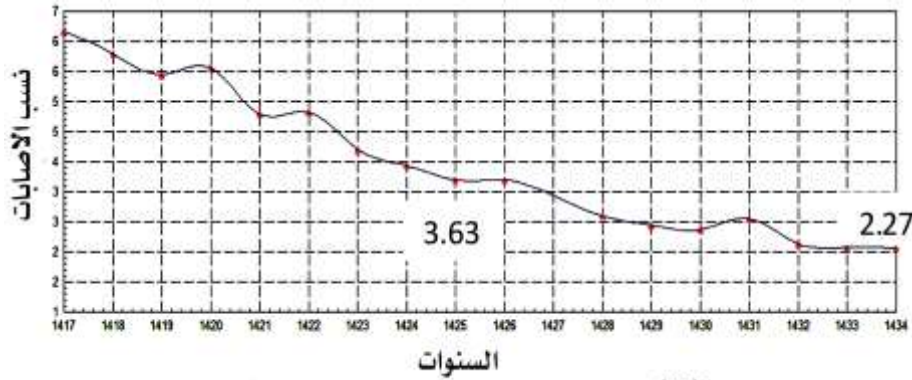
بشكل مماثل لما سبق يمكن تقدير العنصر الدوري، وتحت نفس الفرضيات نلاحظ أن الدالة المقدرة للمركبة الدورية لها الشكل الآتي:

$$\hat{p}(t) = 2.2 \times 10^{-6} + 1.351 \cos(1.257t - 3.172)$$

وبالتالي يصبح لمقدر الطور العشوائي المولد للسلسلة الزمنية الصيغة الرياضية الآتية:

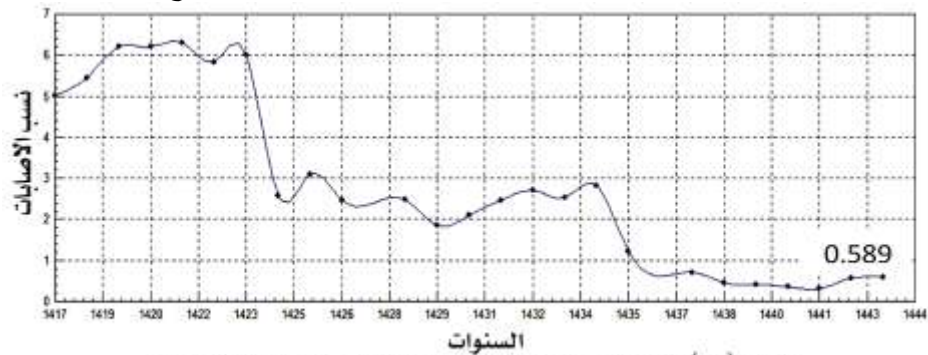
$$\hat{\zeta}_t = 1.61X^3 - 0.67X^2 + 942.04X - 438548.6 + 2.2 \times 10^{-6} + 1.35 \cos(1.257t - 3.172)$$

ومن ثم يصبح لأحد المسارات المقدرة للطور العشوائي المولد للسلسلة الزمنية قيد البحث بالشكل (12):



شكل (12) أحد المسارات المقدرة لنسبة الحوادث ذات الاصابة

وكما سبق سنستعرض أحد المسارات المقدرة للطور العشوائي المولد للسلسلة الزمنية في الشكل (13) نلاحظ أن القيم المقدرة ما زالت تقع ضمن التقدير الجيد، وفقاً للطور العشوائي المقدّر الذي قمنا بتعيينه، مما يزيد ثقتنا في استخدامنا لهذا النموذج في التنبؤ على المدى المسموح به للتنبؤ.



شكل (13) النموذج المقدّر للتنبؤ بنسبة الحوادث ذات الاصابة

ومما سبق نلاحظ أن نسبة المصابين في الحوادث المرورية يظهر تناقصاً واضحاً على مدى السنوات المقبلة، والذي من المتوقع أن تتضاءل نسبة الإصابة إلى حوالي 0.589 حادثة في حلول عام 1443هـ.

4.4.4 استنتاجات

نخلص إلى أن الحوادث المرورية بمنطقة عسير كانت خلال الفترة الزمنية 1417-1434 تسير بشكل متنامي في معظم الحالات، إلا أن هذا السلوك المتنامي أشار إلى مستقبل غير متفائل في أعداد الحوادث، نتيجة الارتفاع المقدّر إذا استمر الوضع على ما كان عليه في الماضي، إلا أنه متفائل بخصوص أعداد ونسبة الإصابات في الحوادث، نتيجة لتوفر وسائل الأمن والسلامة، والذي تجلّى من خلال الصيغ المثلثة لمقدرات الاتجاه العام لكل من السلسلة الزمنية لأعداد الحوادث، والإصابات وكذا نسبة الإصابة، كما أظهرت النتائج وجود عوامل لها دور فاعل في زيادة أعداد الحوادث منها عدم الوعي بأخطار الحوادث، كما أدى توفر وسائل الحماية، وتطور المركبات إلى تناقص أعداد ونسبة الإصابة بشكل ملحوظ.

5. المراجع

1. الإحصاء السنوي للفترة (1417-1434). الإدارة العامة للمرور، منطقة عسير.
2. عبد الحميد البلداوي. (1997): الإحصاء للعلوم الإدارية والتطبيقية. ط1، دار الشروق، عمان.الأردن.
3. وليد دهان الجبوري. (2009): التنبؤ بمستوى التضخم في أسعار المستهلك الشهري في العراق باستخدام السلاسل الزمنية ثنائية المتغيرات. رسالة ماجستير في الإحصاء، جامعة المستنصرية.
4. الشرق الأوسط. (2015): جريدة العرب الدولية، العدد (7981)، المملكة العربية السعودية، طبعة السعودية.
5. سمير مصطفى شعراوي (2005): مقدمة في التحليل الحديث للسلاسل الزمنية. جامعة الملك عبد العزيز، جدة: السعودية.
6. سعدية عبد الكريم طعمة. (2012): استخدام السلاسل الزمنية لتنبؤ بأعداد المصابين بالأورام الخبيثة في محافظة الأنبار. مجلة جامعة الأنبار للعلوم الاقتصادية والإدارية، المجلد 4 العدد 8.
7. حميد العكلة. (2009): تحليل إحصائي لطلاب دراسات الماجستير في الجامعات السورية بين (1991-2006). مجلة جامعة دمشق، المجلد 25، العدد 1، علوم أساسية، سوريا.

8. والتر فاندل. ترجمة عبدالمرضي حامد عزام، أحمد حسين هارون.(1992): السلاسل الزمنية من الوجهة التطبيقية ونماذج بوكس-جيكز - ، دار المريخ، الرياض، السعودية.
9. عصام كوثر وآخرون.(2013): الآثار المترتبة على الحوادث المرورية وسبل تقليصها تاريخ استرجاع 24/2/2014.نشر بموقع: <http://www.al-jazirahonline.com/news/2014/20140224/13208>
10. انتصار أبوتله محمد.(2010): استخدام تحليل السلاسل الزمنية لبناء نماذج الحركة لولاية الخرطوم. رسالة ماجستير في الإحصاء، كلية العلوم- جامعة السودان للعلوم والتكنولوجيا.
11. محمد عبدالعال النعيمي ومحمد حبيب الشاروط.(2000): استخدام التدخل في السلاسل الزمنية لتقدير عدد المصابين بالأورام الخبيثة. مجلة القادسية للعلوم الإدارية والاقتصادية، المجلد 3 العدد 1. العراق.
12. Box G.E.P, and Jenkins, G.M .1976.Time Series Analysis Forecasting and Control (Rev. Ed.). San Francisco: Holden-Day, Inc. First Ed., 19.
13. Brock Well , P.J. and Davis, R.A .1991. Time Series Theory and Methods, 2nd ed , Springer Verlag New York Inc , New York .
14. Clive, W. and Granger .2004.Time Series Analysis Cointegration , and Applications. Paper 2004-02 University of California, San Diego.
15. George C. and Canavos, C .1984. Applied probability and Statistical Methods. Little , Brown and Company-Boston, Toronto 608.
16. Granger C.W.J .1977. Comment on Relationships-and the Lack Thereok- Between Economic Time Series with Special Reference to Money and Interest Rates. by D.A. Pierce. Journal of the American Statistical Association, 72, 357 (March),22-23.
18. Hamilton, J. and James .1994. Time Series Analysis, Princeton U. press, Princeton, NJ ISBN 0-691-04289-6[SJS].
19. Kirchssner G. and Wolters J .2007. Introduction to Modern Time Series Analysis. SPRINGER-Verlag, Berlin Heidelberg.
20. Larsen, R. J. and Marx, M.L .1993. An introduction to Mathematical Statistics and its Applications. Prentice-Hall International, Inc. USA596.
21. Maestas, C. and Robert, R .2000. Modeling Volatility in Political Time Series. Electoral Studies 19:95–110.
22. Mosteller, F., and Tukey, J.W .1977. Data Analysis and Regressions :A Second Course in Statistics. Reading, MA: Addison-Wesley.
23. Murray, R.and Spiegel .2003.Theory and problems of statistics.Mc GRAW Hill Edition, New York.
24. Quenouille, N.H .1957. Analysis of Multiple Time Series. New York, Hafner Publishing Co.
- Shumway, R.H. and Stoffer, D.S .2006. Time Series Analysis and Its Applications. SPRINGER, New York.

Statistical Indicators of Traffic Accidents Using Time Series Analysis in Asir (KSA) during (1417-1434) Hijri
Mohammed M.A. Almazah⁽¹⁾ and Esmail H. A. Alsabri⁽²⁾

Mathematics Dept., College of sciences and Arts (Muhiyl), King Khalid Unvi., KSA; Ibb Unvi., Yemen.

ABSTRACT

Traffic accidents constitute one of the biggest safety threats facing community and have a big impact on its material and human resources. This research provides a time-based statistical analysis of the number of traffic accidents in Asir Region during 1417-1434 hijri (equivalent to 1996-2013), with a focus on determining appropriate probability distributions of the data. It then gives an estimate of the random phase generating the time series in order to reveal the current condition of accidents number, predict future expectations, and anticipate the future behavior of these phases based on a prediction model. The scatter diagram of the time series has revealed a vibrating movement somehow close to harmonic vibrations which represent a cyclical compound and also lack of seasonal compound. It has also been revealed that the appropriate model to display the time series data is the aggregation model. The general trend compound came in the form of non-linear models: that is, the prediction model is non-linear. The results of this research can be considered as an assessment of an important period of time in the history of the rate of traffic accidents and casualties in the region. It has been observed that the number of traffic accidents and casualties has been growing, and is expected to increase to about 66173 incidents by 1443 hijri (2021), despite the expected decrease in the rate of traffic casualties. In addition, the study has pointed out factors which have an active and influential role in traffic accidents rates. It is hoped that this study- in its both theoretical and practical sides- would encourage researchers to apply various predictive models to discuss this subject.

Key Words: Aggregation model, Casualties rate, Harmonic oscillations

Nonlinear Time Series, Traffic accidents.