Journal of Economics, Finance and Management Studies

ISSN(print): 2644-0490, ISSN(online): 2644-0504

Volume 4 Issue 02 February 2021

Article DOI: 10.47191/jefms/v4-i2-04, Impact Factor: 6.228

Page No.- 89-98

Estimation of a Tax Compliance Index with Specific Application to Pay-As-You-Earn

Cathrine T Koloane[1], Mangalani P Makananisa[2]

[1]Masters in Business Leadership in the Graduate School of Business Leadership, University of South Africa, South Africa
Masters Student in the College of Agriculture, Engineering and Science at the University of KwaZulu-Natal, South Africa
[2]MSc in Statistics at the University of South Africa, South Africa

ABSTRACT: This article provides a composite index for Pay-As-You-Earn (PAYE) tax using Principal Components Analysis (PCA). The study uses time series from April 2012 to March 2020 (using monthly data) for the ratios derived from the four compliance measures namely, payments on time, registration on time, filing on time and accurate declarations. The index is computed using the weights of the four derived principal components. According to the model results, the PAYE tax compliance index averages around 75.0% for the period, with the lowest value of 72.3% in 2013/14 and the highest value of 77.1% achieved in 2018/19. There is a clear upward trend, indicating improving levels of compliance in PAYE. Similarly, setting the baseline index of 100 i.e. assuming 100% compliance for 2012/13, results in PAYE tax compliance index averaging around 101.6% for the period, with the lowest value of 97.72% in 2013/14 and the highest value of 104.26% achieved in 2018/19. The study recommends this methodology to be applied to all the tax products and that the overall tax compliance index be computed. This will assist tax authorities all over the world to actively monitor tax compliance levels and institute timeous corrective measures in order to address non-compliance and ultimately maximise PAYE revenue collections. Moreover, this study also serve as a base for many of the future tax compliance indices studies.

KEYWORDS: PAYE Tax Compliance Index, Pay-As-You-Earn, Principal Components Analysis, Normalisation, Standardisation

I. INTRODUCTION

Governments all over the world depend on money collected by tax authorities to better service their countries and citizens. The monies collected by tax authorities is especially critical in developing countries such as South Africa as world class infrastructure such as hospitals, schools, roads etc. still needs to be built and tax revenue serves as a vital source of funding. According to STATSSA (2016), taxes contribute 87% to the total South African government income. This highlights how dependent government is on tax.

Tax compliance means that taxpayers have met their legal obligations under the tax laws and includes ensuring that all tax returns are submitted and tax liabilities paid (SAIT, 2020). The higher the tax compliance levels, the more the government income. The OECD (Forum on Tax Administration, 2004) defines monitoring compliance in four broad categories namely: registration in the system, timely filing or lodgement of requisite taxation information, reporting of complete and accurate information and payment of taxation obligations on time. Compliance relates to the extent to which a taxpayer meets these obligations. If a taxpayer fails to meet any of the obligations then they may be considered to be non-compliant.

The need to monitor overall tax compliance in all the four broad categories has necessitated the development of a composite measure that will serve as a trigger for the tax authorities to investigate adverse trends in tax compliance not only for a particular tax type but for all tax types combined. It is often easier to interpret composite indicators than to identify common trends across many separate indicators and composite indicators are useful in benchmarking country performance (Saltelli, 2007). This study proposes a methodology to compute the PAYE tax compliance index using PCA. The remaining part of the paper is organized as follows: A review of the literature is presented in Section 2. Section 3 discusses the research methodology. Section 4 outlines the data analysis and model results. The paper ends with a conclusion in Section 5.

II. LITERATURE REVIEW

An index is a composite measure of variables, or a way of measuring a construct. It can also be defined as an accumulation of scores from a variety of individual items (Crossman, 2019). There are three steps that can be easily distinguished in index construction, namely: identification of variables, investigating the empirical relationship of these variables and combining these variables into an index and finally validating the index. In literature, several methods are used for constructing indices such as Principal Component Analysis (PCA), Exploratory Factor Analysis (EFA), Partial Least Squares (PLS) regression, Reduced Rank Regression (RRR) and the less complicated additive method. Additive methods can be geometric or arithmetic.

EFA and PCA use the multi-variability between items to derive a new single construct measure. In addition, these techniques provide solutions for assigning different weights to items through the calculation of factor/loading scores. The weights typically present an indication of the relative 'importance' or extent of contribution of specific items towards the final index scores. PLS regression is similar to both PCA and EFA in that predictors are reduced to a smaller set of uncorrelated components in which least squares regression will be performed. RRR model is a multivariate regression model with a coefficient matrix with reduced rank and involves calculating eigenvalues and eigenvectors in determining the index.

de Senna et al. (2019) use PCA to construct the water poverty index by determining the weights of the sub-indexes: resource, access, capacity, use, and environment of the Seridó river basin in the north-east region of Brazil. Five principal components were generated from the data and only components whose eigenvalues were higher than 0.7 were retained. This resulted in three principal components which were used to compose the sub-indexes' weight. An aggregation method suggested by Pérez-Foguet and Giné Garriga (2011) was used to determine the final weights for each sub-index. The overall water poverty index was determined by rescheduling the weights so that the sum of the weights equals one. Several indices were also derived using PCA. SAARF (2009) developed a Living Standard Measure (LSM) to provide a measure of living standard by using criteria such as degree of urbanisation and ownership of cars and major appliances. The variables were selected from a pre-developed questionnaire that was used as part of South African Advertising Research Foundation (SAARF)'s All Media and Products Surveys (AMPS). The variables included the possession of a range of durable household articles, access to reticulated water and electricity supplies, use of the various media, wide range of demographics, including population group, income and education. The index was developed using an index score which was derived based on the loadings of the first factor only. Similarly African Response (2006) developed a well-being index using the first principal component. The index was created in order to measure the socio-economic status or well-being of individuals.

The United Nations developed the Human Development Index (HDI) in order to assess the development of a country through people and their capabilities. The HDI is a summary measure of average achievement in key dimensions of human development namely, long and healthy life, being knowledgeable and having a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions. The indices are normalised by setting maximum and minimum values in order to transform the indicators expressed in different units into indices between 0 and 1. The dimension indices are then calculated as:

$$Dimension\ Index = \frac{actual\ value - minimum\ value}{maximum\ value - minimum\ value} \quad (1)$$

The healthy lifestyle dimension is assessed by life expectancy at birth, the education dimension is measured by the mean of years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age. The standard of living dimension is measured by Gross National Income (GNI) per capita. The HDI uses the logarithm of income to reflect the diminishing importance of income with increasing GNI. The scores for the three HDI dimension indices are then aggregated into a composite index using geometric mean, as shown in Equation 2 below.

$$HDI = (I_{health} . I_{education} . I_{income})^{1/3} \quad (2)$$

Similarly, the Gender Inequality Index (GII), which assesses gender based disadvantage in reproductive health, empowerment and the labour market, is derived using the geometric and harmonic means (United Nations Development Programme, 2019).

Developing a composite index is a challenging and daunting task, full of pitfalls such as the availability of data and the choice of indicators, choice of normalisation and standardisation techniques in order to ensure comparability and aggregation of indicators (weighting and aggregation). Despite these challenges, composite indices are widely used by several international organisations for measuring economic, environmental and social phenomena and, therefore, they provide an extremely useful tool in policy analysis and public communication (OECD, 2008).

III. RESEARCH METHODOLOGY

Different models exist for different purposes depending on the nature of problems to be solved and the availability of data. The study can either fall under supervised, semi-supervised or unsupervised method. The supervised methods are the most popular methods for predictions due to the availability of both explanatory variables and output variable or so-called supervision variable. However, not all problems contains the supervision variable to assess the performance of the model built. In this instance, the unsupervised model might be the best option for data visualisation and grouping of the variables or finding relationships between the variables themselves. However, the two can support each other in coming up with a desirable solution with respect to the trimming of data and/or reliable parameters (for sensitivity analysis).

The current study uses the unsupervised method Principal Components Analysis (PCA) to provide insights on tax compliance trends in South African Pay as You Earn (PAYE). The PCA is concerned with explaining the variance-covariance structures of variables. It is a multivariate statistical technique that linearly transforms a large set of correlated variables into a smaller set of uncorrelated variables (linear combination of these variables), termed Principal Components (PC's) that account for most of the variation in the original set of variables (Johnson and Wichern, 2007).

Let X be the unsupervised *n* x *p* data matrix containing variables $X_{1,}X_2, \dots, X_p$ where each of the n observations form part of the p-dimensional space. PCA will seek a smaller number of dimensions M that explain the entire p-dimensional space where M<p, p represent the number of variables, M the reduced number of variables and *n* the number or observations (or rows) in the data matrix X presented in Table 1 below.

Table 1.The n x p data matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & & x_{nk} & & x_{np} \end{bmatrix}$$

PCA finds a lower dimensional representation of the n x p data matrix that captures most of the information. The dimensions produced by PCA is a linear combination of the p features. The i^{th} principal component (PC_i) is then a linear combination on the p-dimensional space represented by Equation 3 below.

$$PC_i = \mathrm{a}_{1i}X_1 + \mathrm{a}_{2i}X_2 + \cdots + \mathrm{a}_{pi}X_p, \quad i = 1,2,3 \dots \qquad (3)$$

where a_{1i} are loadings/coefficients of the principal component PC_i.

The first principal component (PC_1) will then be represented by the linear combination of variables as shown in equation 4.

$$PC_1 = \mathrm{a}_{11}X_1 + a_{21}X_2 + \cdots + \mathrm{a}_{p1}X_p \qquad (4)$$

The objective is to maximise equation 5, subjected to $\sum_{j=1}^{p} a_{j1}^2 = 1$.

$$\underset{a_{11},\dots,a_{p1}}{Max} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \mathrm{a}_{j1} x_{ij} \right)^2 \right\} \qquad (5)$$

The i^{th} principal components (PC_i) from the original data might have a larger variance compared to the others. However, to perform PCA on the *n* x *p* data matrix (X), the data needs to be normalised/standardised to have mean zero and variance of one such that $\sum_{j=1}^{p} a_{ji}^2 = 1$. This is done for a smoother model, to eliminate the effects of different scales used in the variables and for clear interpretation of the results thereof (James et al., 2013).

Instead of working on the original data set, the standardised data is used for the analysis. Let $z_1, z_2, \dots, z_n$ be the standardised observations with covariance matrix R derived from data matrix, the sample i^{th} PC is shown in equation 6 (Johnson and Wichern, 2007).

$$\hat{y}_i = \hat{e}'_i z = \hat{e}_{i1}z_1 + \hat{e}_{i2}z_2 + \cdots + \hat{e}_{ip}z_p \qquad (6)$$

where $(\hat{\lambda}_i, \hat{e}_i)$ represent the i^{th} eigenvalue and vector pair of R-matrix correlation $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$, $var(\hat{y}_i) = \hat{\lambda}_i,\ i = 1,2,\dots,p$ and $cov(\hat{y}_i, \hat{y}_k) = 0, \quad i \neq k$.

The eigenvectors of the covariance matrix are the directions of the axes where there is the most variance (most information) and that we call Principal Components. Eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each PC. By ranking eigenvectors in order of their eigenvalues, highest to lowest, principal components will be ordered in terms of their descending significance. Moreover, the total standardised sample variance is given by equation 7:

$$var_T(\hat{y}) = tr(\boldsymbol{R}) = p = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p \qquad (7)$$

and the correlation between $\hat{y}_i z_k$ and $\hat{y}_i z_k$ is given by equation 8.

$$r_{\hat{y}_i z_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i},\quad i,k = 1,2,\dots,p \qquad (8)$$

The proportion of the variation explained (PVE) by the i^{th} sample principal component is then given by equation 9.

$$PVE = \frac{\hat{\lambda}_i}{p},\ i = 1,2,\dots,p \qquad (9)$$

Figure 1 is a brief description of each step followed when constructing a PAYE composite index using PCA.

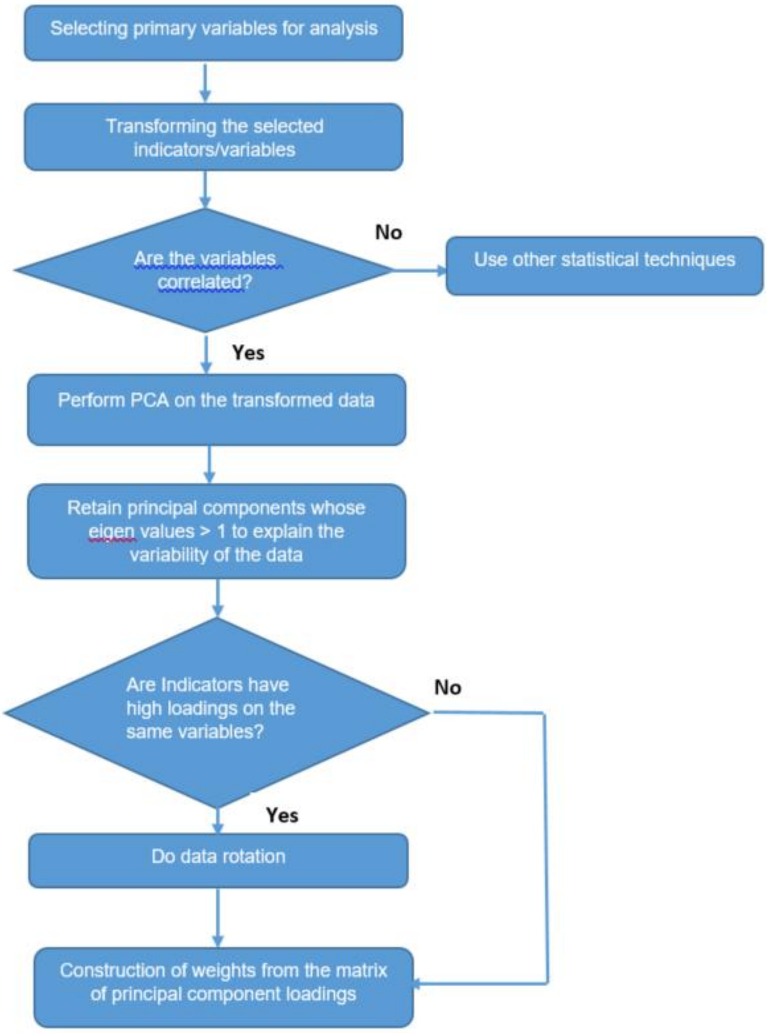

Figure 1. Steps to construct composite index by Principal Component Analysis

To compute a composite index from several principal components, the study adopts a methodology used by de Senna et al. (2019) shown in equation 10 and 11.

$$I = \sum_{i=1}^{p} w_i x_i \qquad (10)$$

where I represents the composite index, w_i represent the final weights and $\sum_{i=1}^{p} w_i = 1$. The weights for each sub-index were aggregated using Equation 11 below.

$$w_i = \sum_{k=1}^{n} \left(a_{k,i} \frac{\sqrt{\lambda_k}}{\sum_{j=1}^{n} \sqrt{\lambda_j}} \right) \qquad (11)$$

where w_i is the final weight used for sub-index *i*, k is the number of PC's, $a_{k,i}$ are the self-vectors that vary from 1 to k and from 1 to *i*, λ_k are the self-values of the principal component *k* and $\sum_{j=1}^{n} \sqrt{\lambda_j}$ is the sum of the j-adopted self-values after the selection criteria application. The weights vary between 0 and 1.

Finally, the compliance index is computed by selecting the baseline index at time t=0 and fix it to 100. Compliance index values for the subsequent periods are then derived using equation 12.

$$CI_t = b_0 + \left[\left(\frac{I_t}{I_0} - 1 \right) * 100 \right] \qquad (12)$$

where CI_t represent the Compliance Index value at time t, b_0 is the baseline index at time 0, I_t represents the Composite Index at time t and I_0 represents the Composite Index at time 0. Thus, the formula measures the deviation from the base year.

IV. DATA ANALYSIS AND MODEL RESULTS

Sample Data

The PAYE historical data consisting of ratios from the four compliance measures namely, payments on time, registration on time, filing on time and accurate declarations, was collected from April 2012 to March 2020 (96 observations). All the variables are showing some form of seasonality. In addition, registration on time (REGProp) is showing a strong increasing trend over time and filing on time (FILEProp) is showing a decreasing trend over time (Figure 2).

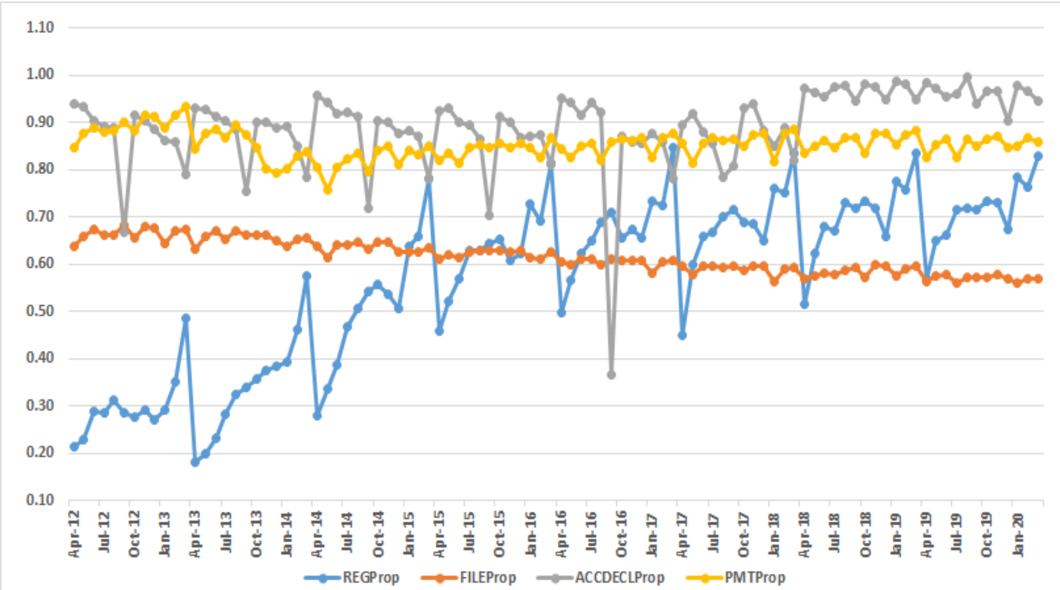

Figure 2. Sample data with registration on time, filing on time, payments on time and accurate declarations

Data Pre-Processing

There was slight variation in the scale and distribution of the ratios compiled from the four compliance measures. As a result, data had to be standardised in order to ensure equal contribution of variables to the analysis and an unbiased outcome. There are many different normalisation and standardisation techniques such as the Min-Max and Distance to a reference methods, however for this paper the Z-score method as described below was adopted.

Data Standardisation/Normalisation

Variables need to be standardised before they are aggregated into composite indicators. Data standardisation ensures that all data sets are on the same scale and thus comparable. PCA is quite sensitive to the variances of the initial variables. If there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with smaller ranges. The most commonly used method for standardisation is the Z-scores. It imposes a normal distribution to the data by converting indicators to a common scale with a mean zero and standard deviation of one. The formula for calculating Z-scores is as follows:

$$Z = \frac{X-\mu}{\sigma} \qquad (13)$$

where X is the individual observations, μ represents the mean of the observations and σ is the standard deviation of the observations. Transformation in Z-scores allows only relative comparisons over time since it is based on the mean and the variance of the indicators at the time of reference.

Table 2 below shows the tabular presentation of the indicators before and after applying data standardisation and normalisation. The standardised data clearly reflects a mean of zero and standard deviation of one.

Table 2. Descriptive Statistics

Table 2a. Before standardisation

Variable	Obs	Mean	Std. Dev.	Min	Max
REGProp	96	.5689515	.1807704	.1810056	.8469161
FILEProp	96	.6160524	.0337447	.5594877	.6838674
ACCDECLProp	96	.8935772	.0842367	.3655172	.9957465
PMTProp	96	.853546	.0292133	.756189	.9339771

Table 2b. After standardisation

Variable	Obs	Mean	Std. Dev.	Min	Max
zreg	96	1.99e-09	1	-2.14607	1.537667
zfile	96	1.94e-10	1	-1.676252	2.009646
zaccdecl	96	3.30e-09	1	-6.268763	1.212882
zpmt	96	-1.58e-09	1	-3.332627	2.753236

Data Analysis and Results

Principal Components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set, followed by the second principal component and the rest of the principal components in a descending order, subject to the constraint that the sum of the squared weights ($a_{11}^2 + a_{12}^2 + a_{13}^2 + \cdots + a_{1n}^2$) is equal to one. Table 3 below shows the eigenvalues or coefficients of each PC in descending order. The eigenvalues of PC_1 and PC_2 are 1.895 and 1.059 respectively. The remaining PC_3 and PC_4 eigenvalues are 0.893 and 0.153 respectively.

Table 3. Principal Components Eigenvalues

Principal Component	Eigenvalue
comp 1	1.895
comp 2	1.059
comp 3	0.893
comp 4	0.153

Figure 3 below is the scree plot for the number of PC's to be retained to explain the data variation. The Kaiser-Meyer-Olkin (KMO) statistical test recommends that all PC's with eigenvalues greater than 1 should be retained. However, for this study all PC's were retained to compute the composite index using all four compliance measures and explains 100% of the variation in the data.

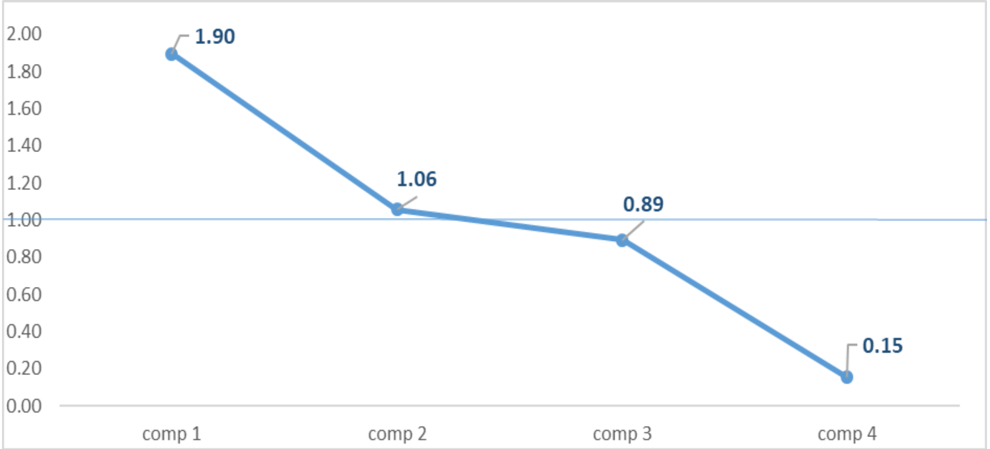

Figure 3. Principal Component Scree plot

The percentage of variation explained by each PC is clearly displayed in Figure 4 below. PC_1 explains 47.4% of the variation and PC_2 explains 26.5% of the variation. The first two PC's cumulatively explain nearly 74% of the variation. PC_4 explain a relatively small variance of 3.8%, thus even if it can be excluded it won't affect the model results significantly. However, the inclusion of the fourth dimension should not be a problem as the aim is to compute the compliance index using as much relevant data as possible.

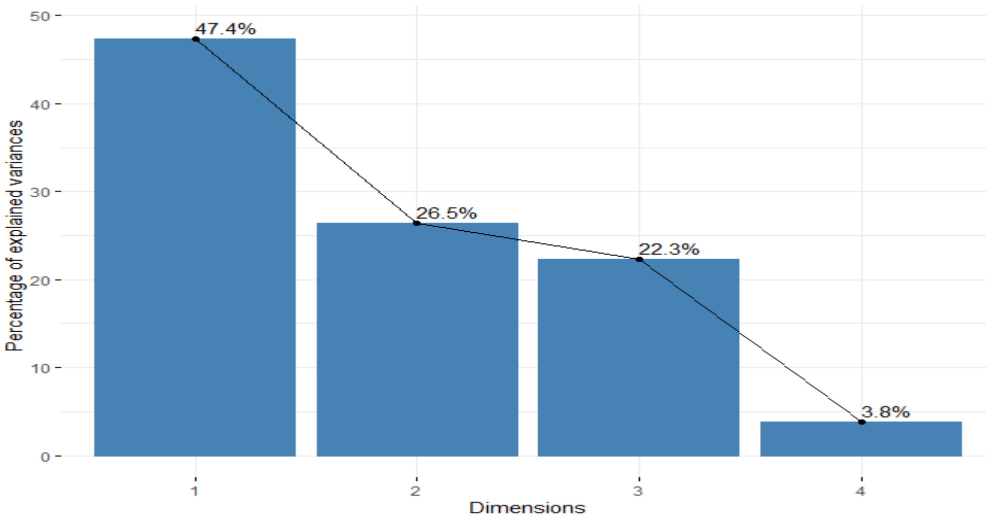

Figure 4. Variance explained by linear combination

One of the interesting analysis of PCA are the variable coordinates presented in table 4 below. It represents the relationship between the variables and the linear combination or principal components. Thus the covariance matrix indicates the weights of the principal components. Within each PC_i where *i*=1,2,3,4, there are eigenvectors by which original variables are multiplied by to compute PC score for any observation.

Table 4. Variable coordinates and final weight

Indicator	Dim.1	Dim.2	Dim.3	Dim.4	VCI	Final Weight
Registration	-0.83	0.49	-0.06	0.25	-2.26	0.14
Filing	0.96	-0.08	-0.04	0.28	4.88	0.30
Payment on time	0.34	0.61	0.71	-0.05	6.55	0.41
Accurate declarations	-0.41	-0.67	0.61	0.10	-2.36	0.15

The first linear combination is shown by equation 14.

$$PC_1 = -0.836reg + 0.96fil + 0.34dec - 0.41pmt \qquad (14)$$

where PC_1 represent the first principal component, *reg* represents registration, *fil* is filling , *dec* is declarations and *pmt* represents the payments variable. Similarly, this can also be derived for the remaining three PC's or dimensions. However, this study adopted the aggregation method used by de Senna et al. (2019) as shown in equation 10 and 11 under the research methodology. Therefore, an additive composite index from PCA results subjected to $\sum_{i=1}^{4} w_i = 1$ is shown in equation 15.

$$I_{PAYE} = 0.14reg + 0.15dec + 0.30fil + 0.41pmt \qquad (15)$$

Since the variables used were standardised to have mean zero and variance one, the variable coordinates are the same as the correlation matrix. However, the correlation highlighting the most contributing indicator/variable for each dimension is visualised in figure 5. The larger the circle and the darker the colour, the stronger the correlation between variables and the PC's/Dimensions. Registration and Filling are highly correlated to the first dimension.

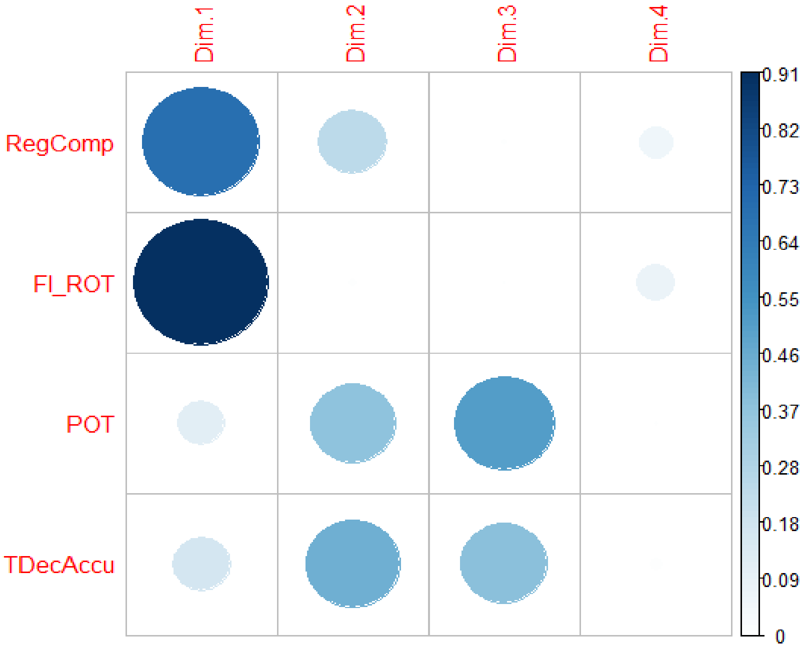

Figure 5. Correlation between variables and the Principal Components

RegCom represent registration, FI_ROT represents filling, POT represents payment on time and TDecAccu represents declarations.

Table 5 below shows the original variables (proportions) and the computed annual PAYE composite index for the period 2012/13 to 2019/20. The average for the PAYE composite index over the eight year period under study is 75.0%.

Table 5. PAYE Composite Index

Year	%Reg	% File	% Dec	% PMT	PAYE Compliance Index
2012/13	30.10%	66.50%	88.40%	89.40%	73.95%
2013/14	33.50%	65.60%	88.60%	84.70%	72.27%
2014/15	52.90%	63.50%	89.50%	82.00%	73.39%
2015/16	63.40%	62.20%	88.20%	84.30%	75.21%
2016/17	67.30%	60.50%	88.90%	85.20%	75.71%
2017/18	68.60%	59.00%	88.40%	85.80%	75.61%
2018/19	70.40%	58.50%	97.00%	86.10%	77.10%
2019/20	71.40%	57.00%	97.90%	85.40%	76.63%

Setting the baseline index of 100 for 2012/13 and thereafter calculating the index for subsequent years, results in an apparent increasing compliance levels as shown by Figure 6 below. Compliance levels have generally been higher than the base period except for 2013/14 (97.72%) and 2014/15 (99.23%). The overall high level of compliance in PAYE is expected as companies/entities are bound by law to supply their employees' third party data to the tax authority and this is reconciled against what the employees personally declare on their annual returns.

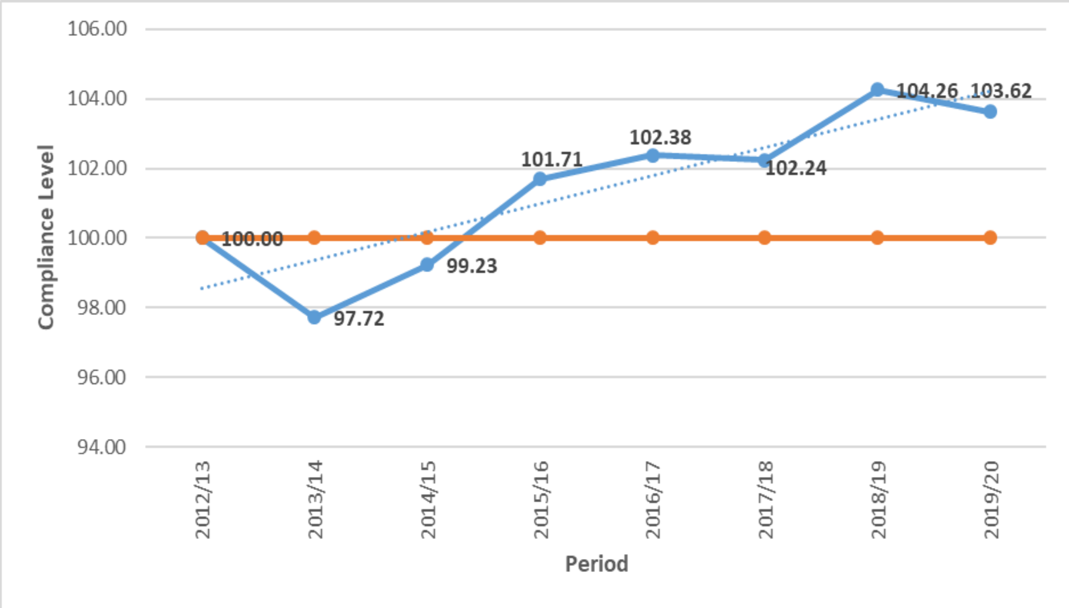

Figure 6.PAYE Compliance Levels

V. CONCLUSION

The results of this study indicates increasing PAYE compliance over the period 2012/13 to 2019/20. The overall high level of compliance in PAYE is expected as companies/entities are bound by law to supply their employees' third party data to the tax authority and this is reconciled against what the employees personally declare on their annual returns.

The methodology adopted in this paper is a momentous stepping stone towards the development of tax compliance indices in South Africa which can be used to drive appropriate management actions to improve compliance. Moreover, once tax compliance indices are developed by most countries in the world, they can be used for benchmarking purposes.

Therefore it is recommended that this methodology be applied to other tax types and ultimately be used to derive the overall tax compliance index.

REFERENCES

1) African Response. 2006. FinScope small business survey report Gauteng 2006. South Africa. Available from :< http://www.finscope.co.za/sme.html>.
2) Crossman, A. 2019. How to Construct an Index for Research. New York: Thought Co. Available from :< https://www.thoughtco.com/index-for-research-3026543>.
3) de Senna, L.D., Maia, A.G. and de Medeiros, J.D.F.2019. 'The use of principal component analysis for the construction of the Water Poverty Index', Brazilian Journal of Water Resources, 24(19):1-14.
4) Forum on Tax Administration. 2004. Guidance Note. Compliance Risk Management: Managing and Improving Tax Compliance. London: Centre for Tax Policy and Administration, OECD.
5) James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. An introduction to statistical learning. 1st Edition, Springer. USA.
6) Johnson, R.A. and Wichern, D.W. 2007. Applied multivariate statistical analysis. 6th Edition, Pearson Prentice Hall. USA
7) OECD. 2008. Handbook on Constructing Composite Indicators. Methodology and user guide. Paris: OECD Publications.
8) Perez-Foguet, A. and Gin^{é} Garriga, R.G.2011. 'Analyzing water poverty in basins', Water Resources Management, 25(14):3595-3612.

9) SAARF. 2009. The SAARF AMPS Living Standards Measure (LSM). South Africa Available from: <http://www.saarf.co.za/AMPS/technicalreport-2009A/data%20files/Technical/21+22.pdf >.
10) SAIT. 2020. Tax Compliance. South Africa: Pretoria. Available from :< https://www.thesait.org.za/news/524096/Tax-Compliance.htm>.
11) Saltelli A. 2007. 'Composite indicators between analysis and advocacy', Social Indicators Research, 81:65-77.
12) STATSSA. 2016. How important is your tax money? South Africa: Pretoria. Available from :< http://www.statssa.gov.za/?p=9155>.
13) United Nations Development Programme.2019. Human Development Reports. Human Development Index (HDI). Available from :< http://hdr.undp.org/en/content/human-development-index-hdi>.