



UNIVERSITY OF NAIROBI

**Modelling the Relationship Between GDP per Capita
and the Informal Economy in Kenya: A Multiple Causes
and Multiple Indicators Approach**

BY

Cynthia Thinwa

I56/32988/2019

A Thesis Submitted to the Department of Mathematics for Examination in Partial
Fulfillment of the Requirements for the Award of Degree of Master of Science in
Social Statistics of the University of Nairobi

November 2021

Abstract

A lot of research around the informal sector examines it from the employment perspective, with a key problem being the difficulty in estimating the true size of the informal sector. Other research focuses on studying this sector as it appears in urban settings; in this context, the challenges that they face as they conduct business are studied. However, their contribution to GDP is an ongoing knowledge gap, with various methods proposed to estimate the size of the sector and the contribution to GDP that it makes. Pursuing this line of thought will enable policy makers to change the narrative from only looking at its expansion in terms of employment, to quantifying its value to the economy, in order to investigate if current interventions such as group credit have made an impact on the production of this sector.

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.



Signature

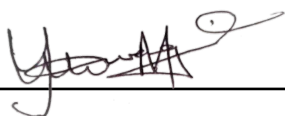
22/11/2021

Date

CYNTHIA THINWA

Reg No. I56/32988/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.



Signature

26/11/2021

Date

Dr John Ndiritu
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: jndiritu@uonbi.ac.ke

Dedication

This project is dedicated to those working in the informal sector, despite its many challenges. It is my hope that the findings of this paper can help the Kenyan government and representatives from the informal sector design actionable policy that involves informal sector workers at the grassroots, removes chronic stumbling blocks that they face and provides material benefit to them.

List of Abbreviations

AIC :	Akaike Information Criterion
CSP Services :	Community, Social & Personal Services
GDP :	Gross Domestic Product
ILO :	International Labour Organization
KSH :	Kenya shillings
MIMIC :	Multiple Indicators Multiple Causes
RMSEA :	Root Mean Square Error of Approximation
RMSR :	Root Mean Square Residual

Contents

Abstract	ii
Declaration and Approval	iv
Dedication	vii
List of Abbreviations	viii
Figures and Tables	xi
Acknowledgments	xiii
1 Introduction	1
1.1 Informality in Today’s World	2
1.2 Problem Statement	4
1.3 Research Objectives	5
1.3.1 General Research Objective.....	5
1.3.2 Specific Research Objectives	6
2 Literature Review	7
2.1 Theoretical Literature Review	7
2.1.1 Conceptualizing The Informal Sector	7
2.1.2 Structural Equation Models	9
2.2 Empirical Literature Review.....	9
2.2.1 Multiple Indicators, Multiple Causes (MIMIC) model.....	9
2.2.2 Panel Data Analysis: Heterogeneity among Individuals	13
2.3 Kenya Specific Review.....	14
3 Methodology	16
3.1 Research Design	16
3.2 Data.....	16
3.3 Empirical Model.....	17
3.3.1 Selected Variables	17
3.3.2 Multiple Indicators Multiple Causes (MIMIC) model	25
3.3.3 Data Validation	29
3.4 Parameter Estimation.....	31
3.4.1 Maximum Likelihood.....	31
3.4.2 Ordinary Least Squares	34
3.4.3 Statistical significance and precision of estimates.....	37
3.5 Goodness of fit tests.....	38
4 Results	41
4.1 Data.....	41

4.1.1	Descriptive statistics	42
4.2	Empirical Model.....	50
4.2.1	Heterogeneity Test Results.....	50
4.3	Parameter Estimation.....	56
4.4	Goodness of Fit.....	59
4.4.1	Cointegration Test Results	59
4.4.2	Statistical Significance of the Model	60
4.4.3	Goodness of fit statistics	61
4.4.4	Model Selection	61
4.5	Model Interpretation.....	62
5	Conclusion	63
5.1	Summary of Results	63
5.2	Comparison & Contrast of Results.....	64
5.3	Conclusions.....	67
5.3.1	Key Conclusions	67
5.3.2	Study Limitations.....	67
5.4	Recommendations	68
5.4.1	Recommendations for Government	68
5.4.2	Recommendations for Policy Makers	68
5.5	Future Research	69
	References	70

Figures and Tables

Figures

Figure 1. Cash Flows Between The Formal and Informal Sector	8
Figure 2. Visual Representations of Problems to Be Solved	20
Figure 3. Total Indirect Tax over Time: Summary Statistics	43
Figure 4. Growth in Total Indirect Tax over Time: Summary Statistics.....	43
Figure 5. Total Public Consumption over Time: Summary Statistics	44
Figure 6. Growth in Total Public Consumption over Time: Summary Statistics.....	44
Figure 7. Proxy Unemployment Rate over Time: Summary Statistics.....	45
Figure 8. Growth in Proxy Unemployment Rate over Time: Summary Statistics	45
Figure 9. Average Time Deposit Interest Rate Declared by Commercial Banks over Time: Summary Statistics.....	46
Figure 10. Growth in Average Time Deposit Interest Rate Declared by Commercial Banks over Time: Summary Statistics	46
Figure 11. Number of Workers in the Informal Sector over Time: Summary Statistics	47
Figure 12. Growth in Number of Workers in the Informal Sector over Time: Summary Statistics.....	47
Figure 13. M1 Money Supply over Time: Summary Statistics.....	48
Figure 14. Growth in M1 Money Supply over Time: Summary Statistics.....	48
Figure 15. GDP at Current Prices over Time: Summary Statistics.....	49
Figure 16. Growth in GDP at Current Prices over Time: Summary Statistics.....	49
Figure 17. Growth in GDP per Capita at Constant Prices over Time: Summary Statistics.....	50
Figure 18. Comparison Between Informal Sector Workers Based on Their Location	51
Figure 19. Summary Statistics for Informal Sector Workers Based on Location.....	51
Figure 20. Comparison Between Informal Sector Workers Based on Their Industry over Time.....	52
Figure 21. Summary Statistics for Informal Sector Workers Based on Industry	52
Figure 22. Reduced Model Based on the Informal Community, Social & Personal Services Industry	62

Tables

Table 1. Multivariate Analysis of Variance applied to the MIMIC model.....	40
Table 2. Panel Datasets Under Analysis.	41
Table 3. Chow's (1960) poolability test results.	50
Table 4. Stationarity Test Results.....	53
Table 5. Normality Test Results for Individual Variables.....	55
Table 6. Statistical Significance and Precision of Estimated Parameters in the Full Model Based on the Informal Trade & Hospitality Industry.	57
Table 7. Statistical Significance and Precision of Estimated Parameters in the Full Model Based on the Informal Community, Social & Personal Services Industry.	57
Table 8. Statistical Significance and Precision of Estimated Parameters in the Reduced Model Based on the Informal Trade & Hospitality Industry.....	58

Table 9. Statistical Significance and Precision of Estimated Parameters in the Reduced Model Based on the Informal Community, Social & Personal Services Industry.	58
Table 10. Cointegration Test Results on Model Residuals.	59
Table 11. Model Statistical Significance Test Results.	60
Table 12. Goodness of Fit Statistics.	61
Table 13. Comparison of Various Author Findings.	65

Acknowledgments

Firstly, I wish to thank the Almighty God for giving me the wisdom, good health and determination that was needed to complete this project.

Secondly, I wish to thank my very supportive family that gave me the space needed to work and participated in many a brainstorming session.

I thank my mentor for keeping me accountable throughout my Masters journey and encouraging me every step of the way.

Last but not least, I wish to thank my supervisor, whose attention to detail, patience in explanation and pursuit of excellence helped transform a very rough idea into the polished work presented in this report.

Cynthia Thinwa

Nairobi, 2021.

1 Introduction

GDP per capita is the total goods and services produced by a nation divided by that particular nation's population. The informal sector in this paper is viewed as the number of people engaged in non-agricultural economic activity for purposes of sustenance, outside of written, formal contracts and government regulation. The aim of this thesis is to determine the nature of the relationship between these two metrics adjusting for other economic indicators. The model of choice is the Multiple Indicators Multiple Causes (MIMIC) model.

The outline of the thesis is therefore as follows:

Chapter 1: A brief introduction to the concept of the informal sector, followed by the problem statement and research objectives.

Chapter 2: A literature review from a theoretical, empirical and localized perspective outlining the concept of the informal sector, methods of estimating its size and explaining how the MIMIC model can solve the research problem.

Chapter 3: A description of the research design, data, model and methods for parameter estimation and evaluation of the model's goodness of fit.

Chapter 4: Results from data analysis and modelling, with selection and analysis of the best model.

Chapter 5: A discussion of the results obtained and comparing with findings from the literature, as well as an assessment of the research study conducted and policy recommendations.

1.1 Informality in Today's World

The informal sector, also described as the shadow economy and the informal economy, has been the subject of research, highlighted in recent times by both global and local media. There are approximately 1 billion workers that form the global informal sector (Benanav, 2019). Therefore, it would be worthwhile studying it and determine the best way to handle it as discerning policymakers.

The informal sector is primarily viewed from a development economics, labour economics and sometimes entrepreneurial lens. The term first came to prominence in Hart's (1985) work where he came across individuals that were neither formally employed nor engaging in subsistence agriculture, yet were conducting economic activities.

Over the years, these individuals have increased in number and the nature of their activities has also grown in complexity. Some of these people have risen to informal employer status, hiring some apprentices and casual labourers; others have been subcontracted work by formal firms; others have chosen to be one-person shops doing everything themselves; others have remained in informal employment; others still, are gig-workers for large technology firms (Alter Chen, 2005).

Global organizations are emerging to become key players that are seeking to engage and profit from informal sector workers through targeting worker subgroups like women and youth, and leveraging aspects of the informal economy such as industry associations, welfare groups and group credit. Wealth for local and offshore formal economies is generated and quality and decent goods and services that these informal workers can use are delivered (Meagher, 2018). This has added further complexity regarding the workings of the informal sector.

It can be tricky to define the informal sector, because some researchers view it as completely separate from the informal sector; others view it in interdependence with the formal sector; others still, view it as not only distinct and separate from the informal sector, but also deliberately choosing to operate illegally, outside of government regulation (Alter Chen, 2005).

In real-life, all three definitions of the informal sector can apply (Alter Chen, 2005), depending on circumstance. For example, a lady can sell from a small stall in a certain neighbourhood for daily sustenance (only informal work). She may have to pay herself first a wage, then send the day's revenue to the owner of the goods and the stall, an office worker (formal work raising capital for informal work). This office worker could choose not to register this business, rationalizing that the business brings revenue in amounts too small to be taxed (choosing to be hidden).

For the purposes of this discussion, criminal "products" such as robbery, fraud, drug-dealing etc. though part of the shadow economy (Georgiou, 2007), will not be studied as part of the informal sector. Illegality is more in terms of corruption that facilitates selling legitimate products, such as paying bribes to city council officers to sell goods in the lucrative Central Business District (Dragsted, 2019).

Due to the unregulated nature of the informal sector, trust is typically the glue that allows for exchanges of value to take place. Business partnerships are formed based on trust, and an informal entrepreneur that violates the trust can be excluded from a particular social network; distrust in the government can also make informal sector players deliberately hide themselves. Criteria particularly in Africa for these trust-based relationships are ethnicity and family-ties (Odera, 2013).

Furthermore, willingness of these informal sector workers to participate in informal economic activity has been a source of much debate. Some people view them as solo entrepreneurs railing against the system, while others view them as unwilling, underemployed workers who are trapped in the informal sector (Benanav, 2019). There is some truth to this, as people who are employers in informal systems dramatically outpace their employees in pay, and most of these workers are labourers or wage workers, not owners; furthermore, men dramatically tend to outearn women in the informal sector, with women sometimes giving their labour without pay (Alter Chen, 2005).

When the informal sector first emerged as a concept, the ILO (1972) only studied urban informal sector workers. In Kenyan urban settings, retrenchments in the 1980s forced formerly employed formal workers to enter the informal sector and earn a living, growing the size of the informal sector in the process. The volumes of goods sold by urban informal workers can be quite large; hawkers in Eldoret, for example, collectively could sell stock worth KSH 45 million in a day (Rotich, 2013).

The Kenyan informal sector is typically referred to as *Jua Kali* because when it started, workers used to work in the hot sun. It is currently integrated with formal enterprises and spans both urban and rural areas of the country. Rural artisans (in groups or alone) are typically engaged with manufacturing low cost and low quality goods, that are then picked by urban traders and sold in urban areas for a profit; however, some of the high-quality furniture sold by formal firms is made by informal wage employees directly employed by those formal firms or sourced from an informal entrepreneur directly then sold at a higher price to Kenya's middle and upper classes (Bigsten et al., 2004).

In the Kenyan case, there appears to be an informal supply chain of sorts, with heterogeneity in the product quality, product price and worker pay. This goes to show that it is important to expand the initial conceptual framework that arose in labour economics and use a more interdisciplinary and data-driven framework.

1.2 Problem Statement

As discussed in the introductory section of this report, there is no general consensus on the conceptual framework and definition of the informal sector. The informal sector in this report would be defined as people not in small-scale agriculture and recognized by the government either working for or owning micro enterprises that do not pay direct taxes. They do not use contracts but form trust-based partnerships to govern how they trade with each other. They can work with people employed by or owning formal enterprises primarily as suppliers (Kenya National Bureau of Statistics, 1973; Alter Chen, 2005; Odera, 2013; Bigsten et al., 2004).

Measurement of the informal sector is important, as it employs a large number of the world's population. Various attempts have been made to measure it with varying degrees of success. Due to the opaque nature of the informal sector and the problem of determining if it is being double-counted or not (Alter Chen, 2005), it is no surprise that determining its size is difficult to do.

Kenya keeps national records of the size of the informal sector (Charmes, 2000) in terms of number of workers (Kenya National Bureau of Statistics, 1973), indicating that it views the informal sector from a strictly employment perspective. However, as discussed earlier, this conceptualization does not explicitly tie their contribution to the national GDP (Charmes, 2000). Another limitation of this reporting style is that it does not account for people informally employed within formal organizations (Charmes, 2012). The final estimates provided by the Kenyan government for each year also differ from report to report, further complicating the process of measuring the size of the informal sector.

Georgiou (2007) notes that methods of measuring the informal sector include use of national surveys targeting households and enterprises, examination of currency or money supply indicators, differences between expenditure and income method, use of indirect measures, use of the Multiple Indicators, Multiple Causes (MIMIC) approach and use of indicators in the labour market. Each of these methods have their pros and cons and without a strong theoretical foundation, it can be difficult to justify selection of one form of measurement over another (Georgiou, 2007). Focusing too much on measurement can also de-emphasise possible causal relationships.

Picking one of these measurement methods, the MIMIC approach, can shift the focus to determining the nature of the relationship between GDP per capita and the size of the informal sector adjusting for other economic indicators. Econometricians need to determine if the MIMIC model can determine this particular relationship given updated time series panel data for the Kenyan economy.

The heterogeneity inherent in the informal sector is also very high, and it can be difficult to determine appropriate groupings in order to develop tailor-made policy. A lot of current research aggregates the informal sector (Georgiou, 2007), therefore it is important to focus on a single country to mitigate heterogeneity and create more generalisable research.

There is not much research having models that use informal sector panel data for a single country; Medina and Schneider (2018) and Charmes (2000) concentrate on world sub-regions, yet the informal sector in Latin America is so different from the one in Africa that they may not be directly comparable (Yusuff, 2011). There is a need to apply MIMIC modelling approach on informal sector panel data provided that there are significant differences between the industries within the sector.

Policy makers need to determine the effect of national GDP on the informal economy, and also estimate how growth in the size of the informal sector has a latent effect on growth of GDP per capita. This will help change the narrative from only looking at its expansion to quantifying its value to the economy. Policymakers also need to consider other factors that do significantly affect the informal economy. Finally, determination of the heterogeneity of the informal sector will help policymakers determine if there is need for creation of tailored programs for each group or not.

1.3 Research Objectives

Given the problem statement outlined above, the general research problem is restricted to determining the relationship between GDP per capita and the informal sector accounting for the heterogeneity within the informal sector. The research question becomes:

What type of relationship does the productivity of an individual (GDP per capita) have with the size of the informal sector in Kenya controlling for other economic indicators and accounting for the heterogeneity within the informal sector?

1.3.1 General Research Objective

- Determine the type of relationship between GDP per capita and the size of the informal sector within the Kenyan economy controlling for confounding economic factors using a MIMIC approach.

1.3.2 Specific Research Objectives

- Determine the effect that growth in the informal sector has on growth in GDP per capita in Kenya.
- Establish the significance of this relationship, controlling for other economic indicators in the Kenyan economy.
- Establish if the government of Kenya should design different policies for different groups within the informal sector or not.

2 Literature Review

2.1 Theoretical Literature Review

2.1.1 Conceptualizing The Informal Sector

It must be appreciated that there is a difficulty in having a global definition and global standards in measuring the informal sector. The informal sector in most of the world regions does make a significant contribution to GDP. It contributed 63.6% in Sub-Saharan Africa, 36.2% in Middle East and North Africa, 30.2% in Asia, 29.2% in Latin America and 19.5% in Transition Countries (Charmes, 2012).

The concept of the informal sector first came to prominence in the 1970s (ILO, 1972) and the 1980s (Hart, 1985). Yusuff (2011) noted that understanding the informal sector needed a critique and synthesis of four perspectives: modernization, dependency, neoliberalism and structuralism. These four perspectives help frame the historical understanding of this sector from the 1960s till present times and these ideologies are currently embedded in the initiatives of non-governmental organizations and their relations with the Global South.

Initially, the informal sector was seen as the backward remnants of traditional, indigenous society; therefore, the aim was reduction of its size through adoption of western ideologies concerned with running a national economy. Then it was viewed as consisting of the poor in society, locked out of participating in national development. Afterwards, it was viewed as a group of entrepreneurs challenging the formal economy, with the potential of replacing the formal sector. The final view was that it was just a parallel to the formal sector that upheld the structure of globalist capitalism and facilitated the driving down of production cost (Yusuff, 2011).

Each of these perspectives has its pros and cons, and two contributed to growing the size of the informal sector. First, the modernization perspective guided policymakers on the privatisation of parastatals which led to mass retrenchments. An unintended consequence of privatisation was that social welfare programs sustained by governments in developing countries were scrapped. Secondly, the structuralism perspective facilitated, ironically enough, the exploitation of wage workers in the informal sector for maximized profit by formal, globalist firms (Yusuff, 2011; Rotich, 2013, Meagher, 2018).

The aforementioned outcomes show the impact of implementing policy based on a faulty understanding of the informal sector. There may be elements of truth in each of the perspectives, but care must be taken to seek a conceptual framework of the informal sector, putting the perspectives of both entrepreneurs and wage workers within the informal sector first while taking a grassroots approach compared to a top-down, trickle economics approach. A conceptual framework visualizing cash flows between the formal and informal sector is as shown:

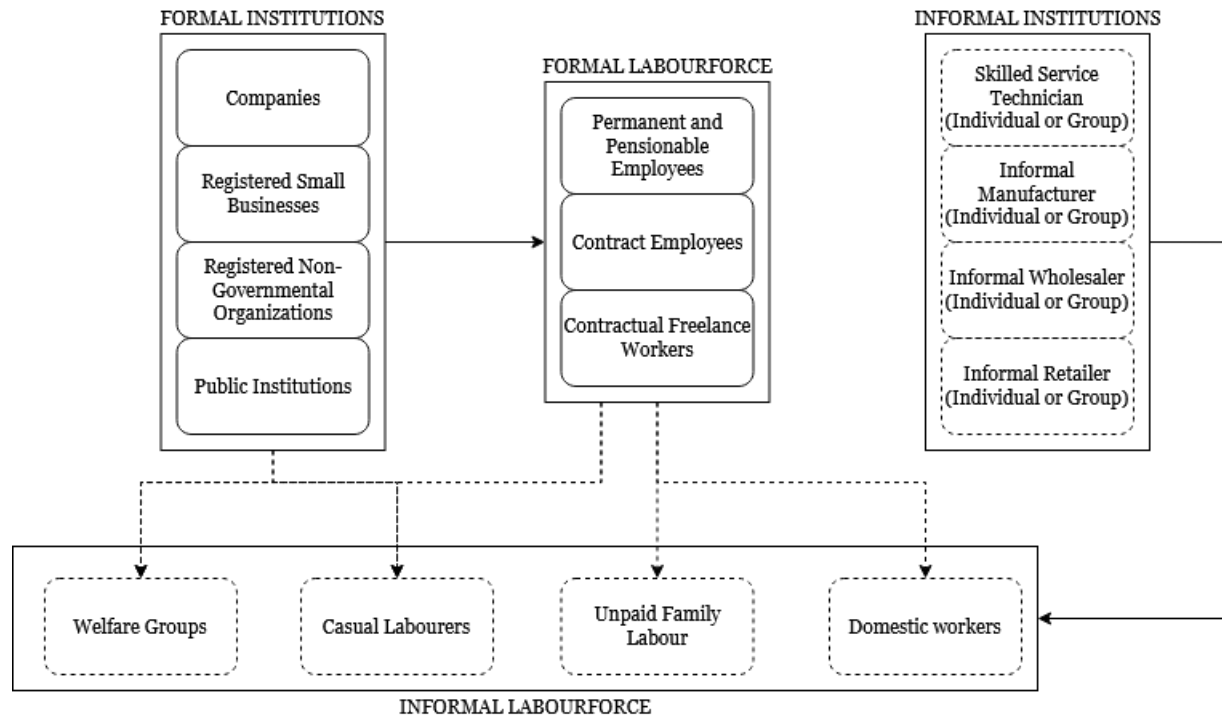


Figure 1. Cash Flows Between The Formal and Informal Sector

What makes the informal sector in Africa different from other parts of the world, is that ethnicity matters a lot, entrepreneurs and employees do informal activities as a means to survival, there are higher barriers to entry and it does have weaker ties with formal firms (Yusuff, 2011; Bigsten et al., 2004; Odera, 2013).

The size of the informal sector according to this report, as recorded by the Kenyan government, would actually be the size of the informal labour force shown in Figure 1 minus unpaid family labour. The informal sector is an interdisciplinary concept with roots in anthropology, economics and applied statistics. Therefore, it is understandable why there is no coherent theoretical economic conceptual framework on the same.

2.1.2 Structural Equation Models

Structural equation models, initially created for use in psychology, were proposed for use in economic settings, first by Goldberger (1972), then by Jöreskog and Goldberger (1975). Frey and Weck-Hanneman (1984) first applied the MIMIC model to relate the relationship between the size of the informal sector and GDP in developed countries, treating it as a latent variable.

Factor analysis represents a hidden factor as a linear combination of observable variables. Structural equation models go one step further beyond this concept by providing indicator variables that the factor can predict; this adds additional dimensionality, changing causal relationships from one-dimensional to two-dimensional (Krishnakumar and Nagar, 2008).

MIMIC model is a special form of a structural equation model.

2.2 Empirical Literature Review

2.2.1 Multiple Indicators, Multiple Causes (MIMIC) model

When it comes to use of models that account for the hidden nature of the informal sector as well as relating it to other economic phenomena, an empirical model is best suited for the task that can not only give a tangible and realistic representation of the size of the informal sector, but also show how it relates to other economic indicators.

$$y = f(\eta)$$

$$\eta = f(x)$$

The MIMIC model is typically specified as shown (Giles, 1999):

Let

η = an index of the size of the hidden economy, the "latent" factor

y = vector of the "indicator" variables

x = vector of the "cause" variables

$$y' = (y_1, y_2, \dots, y_p)$$

$$x' = (x_1, x_2, \dots, x_q)$$

$$\lambda' = (\lambda_1, \lambda_2, \dots, \lambda_p)$$

$$\gamma' = (\gamma_1, \gamma_2, \dots, \gamma_q)$$

$$y = \lambda \eta + \varepsilon$$

$$\eta = \gamma'x + \zeta$$

\therefore

$$y = \lambda(\gamma'x + \zeta) + \varepsilon$$

$$y = \lambda\gamma'x + \lambda\zeta + \varepsilon$$

Let

$$\Pi = \lambda\gamma'$$

$$z = \lambda\zeta + \varepsilon$$

$$Cov(z) = \lambda\lambda'\Psi + \Theta_\varepsilon$$

$$y = \Pi x + z$$

There have been a number of studies leveraging the MIMIC model in various situations. The first study under discussion is that of Giles (1999) where data for New Zealand running for 26 years was studied. Giles (1999) accounted for non-stationarity in the variables used for modelling by making sure that log transformations and differencing were applied appropriately. Furthermore, the MIMIC model also accounted for non-linearity. The informal sector estimate used was

$$\frac{\text{Size of the underground economy}}{\text{Size of GDP}}$$

In Giles' (1999) context, GDP was the real gross domestic product for the period 1982/1983.

According to Giles (1999), the hidden economy had an effect on the ratio of currency to M3 money supply holding male labour force participation rate to unity. Of the cause variables considered in this study, the consumer price index, the ratio of corporate tax to GDP, as well as the ratio of "other" tax to GDP, were the three that were found to have a relatively statistically significant effect on the hidden economy.

However the small sample size and the fact that the effects are interpreted relative to one another made it important to treat the results of model fit with caution when determining the size of the informal sector; that said, the lowest AIC obtained was 49.57 and lowest root mean square residual (RMSR) was 0.04. The best MIMIC model had a RMSR of 0.10 and AIC of 173.77.

Upon applying the currency demand equation to get real-world estimates, Giles (1999) found that the model had a R^2 of 98.2% and the results indicated that the New Zealand informal sector contributed an average of 8.8% in 1981 to real GDP in the long-run. The tax gap due to not accounting for the informal sector was found to be 6.4-10.2% of tax liability (Giles, 1999).

Neither GDP per capita nor growth in GDP per capita were added to Giles' (1999) full model; a differenced form of the logarithm of GDP was used and the informal economy was found to have a positive effect on it.

In another study that used a sample of 158 countries' data collected for 24 years, most of the "cause" variables were found to be significantly affecting the size of the informal economy, which in turn had a statistically significant effect on the labour force participation rate and growth on GDP per capita holding currency to unity (Medina and Schneider, 2018).

Upon reducing the panel data set to strictly developing countries, only trade openness, GDP per capita, unemployment rate, size of government and fiscal freedom among the "cause" variables were found to have an overall statistically significant effect. When specifically comparing MIMIC predicted estimates of the shadow economy with estimates derived from discrepancies in national accounts for some Sub-Saharan countries, the MIMIC model predicted estimates were much smaller (Medina and Schneider, 2018).

In Medina and Schneider's (2018) study, all variables were kept in their original form. GDP per capita was used as a causative variable and growth in GDP was used as an indicator variable. GDP per capita was found to have a negative effect on the informal sector and in turn, the informal sector was found to have a negative effect on growth in GDP per capita. The MIMIC models relating GDP per capita, the informal sector and GDP per capita growth had RMSEA (Root Mean Square Error of Approximation) of 0.055-0.103.

To check model robustness in their model, GDP per capita was omitted and growth in GDP was replaced with night-light intensity; the optimal form of the latter model was found to have a lower RMSEA than the optimal form of the former. The lowest RMSEA score obtained for any of the MIMIC models was 0.01; that MIMIC model involved replacing GDP-related measures with night-light intensity as an indicator variable (Medina and Schneider, 2018).

Medina and Schneider (2018) estimated that, on average, Kenya's informal sector was estimated to contribute 33.2% of its GDP during 1991-2015. In both the Giles (1999) study and the Medina and Schneider (2018) study, the MIMIC model was first applied to get

relative estimates of the informal economy before use of the currency demand equation to form more absolute estimates.

Unlike Giles (1999), Medina and Schneider (2018) did not only use unity restriction on an indicator variable but also used the informal sector estimate's mean and variance obtained from panel data; they also kept the variables in a non-stationary state. The two studies also differ in the choice of data; Giles, 1999 used data only from New Zealand, while Medina and Schneider, 2018 used data from a variety of countries.

Barbosa et al. (2013) used this model to estimate the size of the informal economy in Portugal over a period of 34 years with data collected semi-annually. To ensure comparability, all variables were converted into percentages. The variables were then differenced to obtain their stationary form.

Barbosa et al. (2013) found that unemployment rate and proportion of government subsidies to GDP were statistically significant variables and opted to use only those cause variables for a more optimized model. To get the absolute value of the size of the informal economy, Barbosa et al. (2013) used Schneider's (2005) estimate for Portugal, 1995.

Macias and Cazzavillan (2010) estimated the size of the informal sector in Mexico using data collected annually for 36 years, inspired by the "street vendors" prevalent in Mexico and neighbouring developing countries. They held GDP to a fixed scalar, 1, in some of the models and in other models they held currency to unity; the variables were also kept in their original form.

Macias and Cazzavillan (2010) also used mean estimates of the informal sector at each time t from other sources and scaled them using GDP, just like Giles's (1999) estimate. Inflation, salaries and unemployment were found to be statistically significant cause variables (Macias and Cazzavillan, 2010).

Once Macias and Cazzavillan (2010) did this estimation, they used the currency demand approach as a different benchmark and compared the two time series of absolute values of the informal sector relative to GDP. Just like Medina and Schneider (2018), they opted to keep the variables in their original state without transformation. It is also noteworthy that just like in the case of Giles (1999) and Barbosa et al. (2013), GDP per capita and Growth in GDP per capita are not included in the model.

In the case of all the aforementioned authors, their main goal was to use this model to estimate the absolute size of the informal economy, hence the need to use an additional measurement to get more absolute values of the same. They also maintained use of Giles' (1999) informal sector estimate.

Upon further examination of application of various forms of MIMIC models in the study of the informal sector, some researchers do not use GDP per capita as a cause variable neither do they use growth in GDP per capita as an indicator variable; various forms of money supply and currency are used as indicator variables (Gulzar et al., 2010; Ogbuabor and Malaolu, 2013; Tonuchi et al., 2020).

Other researchers like Njangang et al. (2018) used growth in GDP per capita as a cause variable because they only needed the measurement equation component of the MIMIC model; they found that for their purposes, Generalized Method of Moments worked provided the data was stationary and instruments used were valid.

2.2.2 Panel Data Analysis: Heterogeneity among Individuals

The informal sector, particularly in the case of African countries, unfortunately has few studies that study its heterogeneity; research has therefore been done from an employment perspective to study wage gaps between the formal and informal sector and possible variables affecting them (Nordman et al., 2016, Bargain and Kwenda, 2014).

Upon studying data collected from 6069 workers in Madagascar for a period of 4 years, Pooled Ordinary Least Squares, Fixed Effects Ordinary Least Squares and Quantile Regression models were used to check if gaps in earnings between the formal and informal sector (represented by hourly hours transformed into logarithm form) were affected by worker characteristics, firm characteristics and fixed effects attributable to timing. It was found that pooled ordinary least squares explained most of the variance in the data compared to the fixed effect model; it had the highest R squared statistic (Nordman et al., 2016).

The models used could be applied to this study, not for income comparisons like the aforementioned authors, but strictly to analyze the informal sector as a standalone entity. The fixed effects model discussed by Nordman et al. (2016) was of the form

$$y_{it} = x'_{it}\beta + \gamma I_{it} + \alpha_i + u_{it}$$

where

y_{it} = observation of individual i at time t

x_{it} = vector of k characteristics belonging to individual i at time t

I_{it} = status of a dummy variable for individual i at time t

u_{it} = error in the model

Additionally, the key concern of this study is to determine if the informal sector is homogeneous i.e. the individual-specific effect is not statistically significant, or if it is hetero-

geneous. Therefore the poolability test as proposed by Chow (1960) becomes

$$H_0 : \gamma = 0 \quad \text{vs.} \quad H_1 : \gamma \neq 0$$

resulting in two models relevant to this study:

$$y_{it} = x'_{it}\beta + \gamma I_{it} + \alpha_i + u_{it} \quad (1)$$

$$y_{it} = x'_{it}\beta + \alpha_i + u_{it} \quad (2)$$

Let

$SSE_{(1)}$ = sum of squares of all u_{it} for model (1)

$SSE_{(2)}$ = sum of squares of all u_{it} for model (2)

n = total number of individuals

T = total number of time periods

$df_{(1)}$ = $nT - k$ = degrees of freedom for model (1)

$df_{(2)}$ = $n - k$ = degrees of freedom for model (2)

The test statistic becomes:

$$F_c = \frac{\frac{SSE_{(2)} - SSE_{(1)}}{df_{(2)}}}{\frac{SSE_{(1)}}{df_{(1)}}} \sim F(df_{(2)}, df_{(1)})$$

Reject H_0 if $F_c > F_{\alpha=0.05}(df_{(2)}, df_{(1)})$

2.3 Kenya Specific Review

Kenya has played a leading role around the informal sector. The concept around the informal sector was coined as a result of ILO's (1972) mission to Kenya, combined with Hart's earlier work. Furthermore, Kenya was also instrumental in defining how this sector ought to be measured (Charmes, 2012).

When it was first recorded by the Kenyan government, it focused only on urban areas (Kenya National Bureau of Statistics, 1977), but expanded over time to include rural areas (Kenya National Bureau of Statistics, 1987). The government then decided to focus on industries in the informal sector (Kenya National Bureau of Statistics, 1990).

This reflects the evolution of thought around the informal sector discussed earlier, from viewing it as only a feature of rural-urban migration to realizing that the informal sector in Kenya was more complex than that (Mitullah, 2004).

The figures around the informal sector were kicked off by surveys reported in annual reports published by the Kenya National Bureau of Statistics; however, no further analysis was conducted in these reports. Ouma et al.'s (2007) report kicked off efforts in doing further analysis around the informal sector. Their view of the informal sector included criminal "products" and the key goal of their research was to measure the size of this sector. They used the currency demand approach to do this, assuming constant currency velocity.

Ouma et al.'s (2007) research found that the informal sector contributed 10.51-30.8% to Kenya's GDP. Their model had an adjusted R^2 of 97.7%.

Charmes' (2012), Nchor and Adamec's (2015) and Medina and Schneider's (2018) approach was to analyze Kenya's informal sector within a group. However, it would be valuable to apply the MIMIC model to the Kenyan economy alone, and not within a group.

3 Methodology

3.1 Research Design

The research design will have to be non-experimental because the data is real-world data collected as it occurs (Edmonds and Kennedy, 2016); the nature of the data makes it easier to interpret the findings of this study. Due to the fact that the data was collected on an annual basis in the form of surveys, longitudinal research design is suitable to estimate long-run relationships amongst GDP per capita, the informal sector and growth in GDP per capita as well as determine heterogeneity within the informal sector from location and industry perspectives.

According to Edmonds and Kennedy (2016), non-experimental research conducted through the survey approach faces threats to external, construct and statistical conclusion validity as well as low response rates. To mitigate the effects of low response rates, secondary data from a single source was used. Furthermore, the source, Kenya National Bureau of Statistics, has wide geographic coverage across the country, ensuring that the results obtained are as generalizable as possible, reducing threat to external validity in the process.

3.2 Data

The data source is multiple documents by one author. The Kenya National Bureau of Statistics (1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996a, 1997, 1998, 1999, 2000, 2001, 2002a, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017a, 2018, 2019, 2020) has consistently conducted economic surveys before and after the year Kenya became a republic, 1964. However, due to the fact that the country was building capacity in data collection and statistical analysis, the data collected from 1972 upto 2019 reflects not only evolutions within the data itself, but also evolution in the government agency.

Reported estimates by the government of the working age population during the times when a national census was conducted (Kenya National Bureau of Statistics, 1981; Kenya National Bureau of Statistics, 1996b; Kenya National Bureau of Statistics, 2002b; Kenya National Bureau of Statistics, 2017b) were also collected.

These were used by the author to derive a variable of interest, with the oldest estimate of the working population based on the decade-on-decade growth rate obtained from existing sources.

A panel dataset was then constructed, having a time series of averages for the total number of informal sector workers, as well as other variables of interest such as GDP per Capita at current prices, growth in GDP per capita at constant prices and controls. The time period was 1972-2019.

Economic survey 1973 - Economic Survey 1986 conducted by the Kenya National Bureau of Statistics concentrated on the informal sector workers that were only within the urban areas, then Economic Survey 1987 onwards break down the number of informal sector workers based on their location, in order to determine if they were working in urban or rural areas. Thus, a panel dataset of averages for each estimate i at time t was constructed categorizing the number of informal sector workers by location for the time period of 1983-2019.

Economic survey 1985 onwards, conducted by the Kenya National Bureau of Statistics, categorized informal sector workers by the industries that they worked in. Based on this information, a panel dataset of averages for each estimate i at time t was constructed once more but categorizing the number of informal sector workers by industry for the time period of 1985-2019.

Due to these various breakdowns, there is now more granularity of data on the informal sector in Kenya than before and more insights can be derived from the data. Di Zio et al. (2016) recommend keeping meticulous metadata that explain the changes that the data has undergone, as well as encoding of the various variables under study. Therefore, each economic survey and its metadata was kept as a separate spreadsheet, with a final spreadsheet representing the averaged estimates, ready for analysis.

3.3 Empirical Model

3.3.1 Selected Variables

Based on past empirical studies, annual growth rates in proxy unemployment rate, indirect tax, size of government, and time deposit interest rate were the cause variables selected. The annual growth rate in the size of the informal sector was then identified as the latent variable. Finally, annual growth rates in GDP per capita (at constant prices), GDP (at current prices) and M1 money supply were selected as indicator variables. Growth rates instead of the raw variables were used to enhance interpretability (Klarić, 2011).

Hence, the full model is a 4-1-3 model. The variables were defined as follows:

Proxy Unemployment Rate

Proxy Unemployment was calculated by the author due to unemployment data for Kenya being hard to come by for the 48 years under review. It is the percentage of the working age population that was inactive, unemployed, or in small-scale agriculture and it was calculated using the formula below:

$$\text{Proxy Unemployment} = 1 - \frac{\text{Number of persons engaged}}{\text{Working Age Population as at Latest Census}}$$

The growth rate for year t was calculated as:

$$\text{ProxyUnem} = \frac{\text{Proxy Unemployment}_t - \text{Proxy Unemployment}_{t-1}}{\text{Proxy Unemployment}_{t-1}}$$

Indirect Tax Burden

Total Indirect Tax was obtained as is from the literature, isolated from direct taxes. Direct taxes may be misleading because unemployed people may receive remittances that they do not pay taxes on; some employed people work in the informal sector on a part-time basis and run "side hustles" which they do not pay taxes on. Additionally, there are many temporary labourers in formal firms that also do not pay taxes on their income as they fall below the taxable bracket; some self-employed individuals may chose to write off their incomes as a business expense and reduce their tax obligations. Finally, self-employed informal workers may receive only cash or mobile money transfers in order to reduce direct taxes.

The growth rate for year t was calculated as:

$$\text{ITburden} = \frac{\text{Total Indirect Taxes}_t - \text{Total Indirect Taxes}_{t-1}}{\text{Total Indirect Taxes}_{t-1}}$$

Size of Government

Public Consumption was chosen as the indicator for Size of Government and it was obtained as is from the literature. This quantifies the effect that a large government could have on the informal sector.

The growth rate for year t was calculated as:

$$\text{GovSize} = \frac{\text{Public Consumption}_t - \text{Public Consumption}_{t-1}}{\text{Public Consumption}_{t-1}}$$

Time Deposit Interest Rate

Time Deposit Interest Rate was obtained as is from the literature then averaged by the author; it is the average interest rate declared by commercial banks on short-term deposits i.e. money deposited for 1 year or less. It was calculated as follows:

$$\text{Time Deposit Interest Rate} = \frac{r^{30\text{days}} + r^{3\text{months}} + r^{6\text{months}} + r^{9\text{months}} + r^{12\text{months}}}{5}$$

The growth rate for year t was calculated as:

$$\text{TimeDepIR} = \frac{\text{Time Deposit Interest Rate}_t - \text{Time Deposit Interest Rate}_{t-1}}{\text{Time Deposit Interest Rate}_{t-1}}$$

Size of the Informal Sector

Size of the Informal Sector (abbreviated as *nIS*) was obtained as is from the literature; it is the estimated total number of workers working for the informal sector. It can be categorized from two perspectives: *Location* and *Industry*.

Location

Location (abbreviated as *Loc*) was obtained as is from the literature, forming a 1983-2019 panel dataset; it indicated if the estimated number of informal sector workers in a given year work in urban areas or rural areas. For example, in 1985, the estimated number of informal works was found to be 1000 workers; of these, 300 worked in *rural* areas and 700 worked in *urban* areas.

Industry

Industry (abbreviated as *Ind*) was obtained as is from the literature, forming a 1985-2019 panel dataset; it associated a given estimate of the informal sector with a particular industry. Using the same example of an estimated number of informal sector workers in 1985 being 1000, 200 work in *Manufacturing*, 50 work in *Trade, Hotels and Restaurants*, 175 work in *Community, Social and Personal Services* (CSP Services), 350 work in *Transport and Communications* and the rest worked in *Other industries*.

The problems requiring panel data analysis of the informal sector can be visualized in Figure 2. In line with study objectives, panel data analysis of the 1983-2019 and the 1985-2019 datasets, will be applications of Chow's (1960) poolability test, as discussed earlier. There are therefore two sets of hypotheses to be tested.

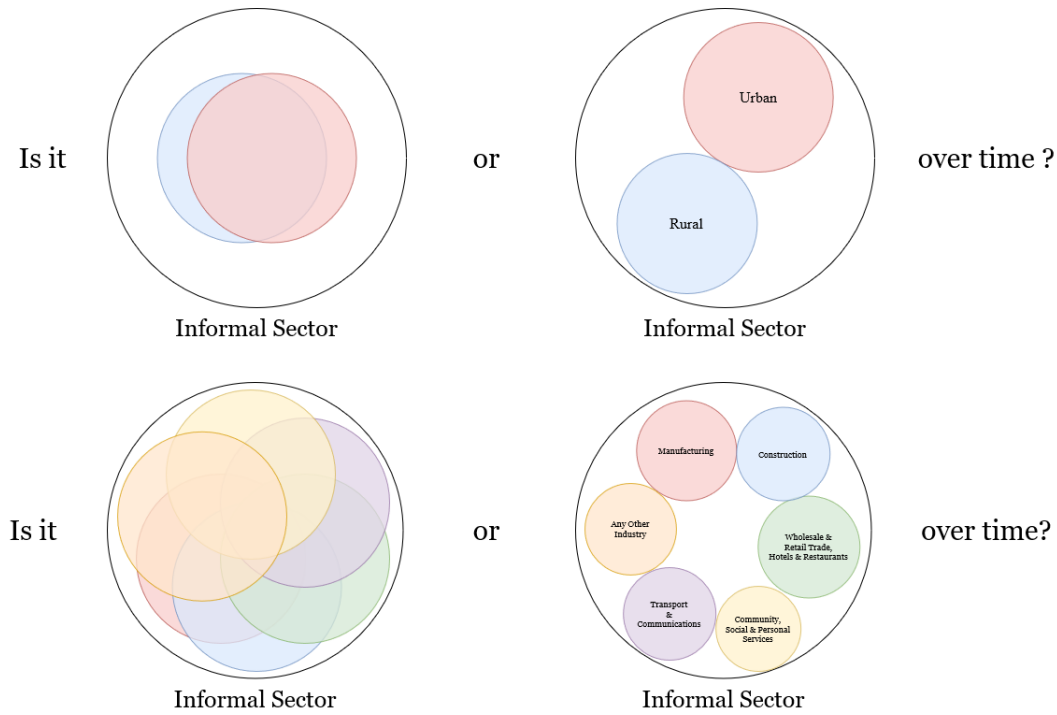


Figure 2. Visual Representations of Problems to Be Solved

1983-2019: Size of Informal Sector by Location

Let

$$Loc \begin{cases} 0 = Urban \\ 1 = Rural \end{cases}$$

$$\eta = \begin{pmatrix} \eta_{1,1} & \eta_{1,2} & \dots & \eta_{1,37} \\ \eta_{2,1} & \eta_{2,2} & \dots & \eta_{2,37} \end{pmatrix}$$

α_i = constant treatment effect attributable to each location grouping

$$\alpha = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \dots & \alpha_{1,37} \\ \alpha_{2,1} & \alpha_{2,2} & \dots & \alpha_{2,37} \end{pmatrix} \quad \text{where} \quad \begin{cases} \alpha_{1,1}, \alpha_{1,2}, \dots, \alpha_{1,37} = \alpha_0 = 0 \\ \alpha_{2,1}, \alpha_{2,2}, \dots, \alpha_{2,37} = \alpha_1 \end{cases}$$

$$D_1 = \begin{cases} 1 \text{ if } Loc = 1 \\ 0 \text{ otherwise} \end{cases}$$

β = treatment effect attributable to location as a whole

u_{it} = random disturbance/innovations in the model

$$\mathbf{u} = \begin{pmatrix} u_{1,1} & u_{1,2} & \dots & u_{1,37} \\ u_{2,1} & u_{2,2} & \dots & u_{2,37} \end{pmatrix}$$

The model under study is

$$\eta = \gamma + D_1\alpha + \beta\eta + \mathbf{u} \quad (3)$$

The hypothesis to check for heterogeneity when the informal sector is grouped by location becomes

H_0 : location treatment effect is not statistically significant i.e. $\alpha_i = 0$

H_1 : location treatment effect is statistically significant i.e. $\alpha_i \neq 0$

According to H_0 , the model becomes

$$\eta = \gamma + \beta\eta + \mathbf{u} \quad (4)$$

with $n - k$ (1) degree of freedom.

However, according to H_1 , the model remains as is, with $nT - k$ (69) degrees of freedom.

Let α level of significance = 0.05

The test statistic then becomes:

$$F_c = \frac{\frac{SSE_{(4)} - SSE_{(3)}}{df_{(4)}}}{\frac{SSE_{(3)}}{df_{(3)}}} \sim F(df_{(4)}, df_{(3)})$$

Reject H_0 if $F_c > F_{0.05}(df_{(4)}, df_{(3)})$

1985-2019: Size of Informal Sector by Industry

Let

$$Ind \left\{ \begin{array}{l} 0 = \text{Any Other Industry} \\ 1 = \text{Manufacturing} \\ 2 = \text{Construction} \\ 3 = \text{Trade, Hotels \& Restaurants} \\ 4 = \text{Community, Social \& Personal Services} \\ 5 = \text{Transport \& Communications} \end{array} \right.$$

$$\eta = \begin{pmatrix} \eta_{1,1} & \eta_{1,2} & \dots & \eta_{1,35} \\ \eta_{2,1} & \eta_{2,2} & \dots & \eta_{2,35} \\ \vdots & \vdots & & \vdots \\ \eta_{6,1} & \eta_{6,2} & \dots & \eta_{6,35} \end{pmatrix}$$

γ = constant treatment effect attributable to overall model intercept

$$\gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \dots & \gamma_{1,35} \\ 0_{2,1} & 0_{2,2} & \dots & 0_{2,35} \\ \vdots & \vdots & & \vdots \\ 0_{6,1} & 0_{6,2} & \dots & 0_{6,35} \end{pmatrix} \quad \text{where } \gamma_{1,1}, \gamma_{1,2}, \dots, \gamma_{1,35} = \gamma$$

α_i = constant treatment effect attributable to each industry grouping

$$\alpha = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \dots & \alpha_{1,35} \\ \alpha_{2,1} & \alpha_{2,2} & \dots & \alpha_{2,35} \\ \vdots & \vdots & & \vdots \\ \alpha_{6,1} & \alpha_{6,2} & \dots & \alpha_{6,35} \end{pmatrix} \quad \text{where } \left\{ \begin{array}{l} \alpha_{1,1}, \alpha_{1,2}, \dots, \alpha_{1,35} = \alpha_0 = 0 \\ \alpha_{2,1}, \alpha_{2,2}, \dots, \alpha_{2,35} = \alpha_1 \\ \alpha_{3,1}, \alpha_{3,2}, \dots, \alpha_{3,35} = \alpha_2 \\ \alpha_{4,1}, \alpha_{4,2}, \dots, \alpha_{4,35} = \alpha_3 \\ \alpha_{5,1}, \alpha_{5,2}, \dots, \alpha_{5,35} = \alpha_4 \\ \alpha_{6,1}, \alpha_{6,2}, \dots, \alpha_{6,35} = \alpha_5 \end{array} \right.$$

$$D_1 = \begin{cases} 1 & \text{if } Ind = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{if } Ind = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{if } Ind = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$D_4 = \begin{cases} 1 & \text{if } Ind = 4 \\ 0 & \text{otherwise} \end{cases}$$

$$D_5 = \begin{cases} 1 & \text{if } Ind = 5 \\ 0 & \text{otherwise} \end{cases}$$

β = treatment effect attributable to industry as a whole

u_{it} = random disturbance/innovations in the model

$$\mathbf{u} = \begin{pmatrix} u_{1,1} & u_{1,2} & \dots & u_{1,35} \\ u_{2,1} & u_{2,2} & \dots & u_{2,35} \\ \vdots & \vdots & & \vdots \\ u_{6,1} & u_{6,2} & \dots & u_{6,35} \end{pmatrix}$$

The model under study is

$$\eta = \gamma + D_1\alpha + D_2\alpha + D_3\alpha + D_4\alpha + D_5\alpha + \beta\eta + \mathbf{u} \quad (5)$$

The hypothesis to check for heterogeneity when the informal sector is grouped by industry becomes

H_0 : industry treatment effect is not statistically significant i.e. $\alpha_i = 0$

H_1 : industry treatment effect is statistically significant i.e. $\alpha_i \neq 0$

According to H_0 , the model becomes

$$\eta = \gamma + \beta\eta + \mathbf{u} \quad (6)$$

with $n - k$ (5) degrees of freedom.

However, according to H_1 , the model remains as is, with $nT - k$ (221) degrees of freedom.

Let α level of significance = 0.05

The test statistic then becomes:

$$F_c = \frac{\frac{SSE_{(6)} - SSE_{(5)}}{df_{(6)}}}{\frac{SSE_{(5)}}{df_{(5)}}} \sim F(df_{(6)}, df_{(5)})$$

Reject H_0 if $F_c > F_{0.05}(df_{(6)}, df_{(5)})$

If the number of informal sector workers is deemed heterogeneous from a location standpoint, the growth rates for year t would be calculated as:

$$Urban_nIS = \frac{No. \text{ in Urban areas}_t - No. \text{ in Urban areas}_{t-1}}{No. \text{ in Urban areas}_{t-1}}$$

$$Rural_nIS = \frac{No. \text{ in Rural areas}_t - No. \text{ in Rural areas}_{t-1}}{No. \text{ in Rural areas}_{t-1}}$$

If the number of informal sector workers is deemed heterogeneous from an industry standpoint, the growth rates for year t would be calculated as:

$$Manufacturing_nIS = \frac{No. \text{ in Manufacturing}_t - No. \text{ in Manufacturing}_{t-1}}{No. \text{ in Manufacturing}_{t-1}}$$

$$Construction_nIS = \frac{No. \text{ in Construction}_t - No. \text{ in Construction}_{t-1}}{No. \text{ in Construction}_{t-1}}$$

$$Trade\&Hospitality_nIS = \frac{No. \text{ in Trade\&Hospitality}_t - No. \text{ in Trade\&Hospitality}_{t-1}}{No. \text{ in Trade\&Hospitality}_{t-1}}$$

$$TransComms_nIS = \frac{No. \text{ in Transport\&Communications}_t - No. \text{ in Transport\&Communications}_{t-1}}{No. \text{ in Transport\&Communications}_{t-1}}$$

$$CSP_nIS = \frac{No. \text{ in Community, Social\&Personal}_t - No. \text{ in Community, Social\&Personal}_{t-1}}{No. \text{ in Community, Social\&Personal}_{t-1}}$$

$$Other_nIS = \frac{No. \text{ in Other industries}_t - No. \text{ in Other industries}_{t-1}}{No. \text{ in Other industries}_{t-1}}$$

GDP per capita

Annual growth in Gross Domestic Product per capita at constant prices (abbreviated as *GDPpc*) was obtained as is from the literature; it is the estimated growth in GDP per capita keeping prices constant over the years.

GDP

Gross Domestic Product at current prices was obtained as is from the literature; it is the estimated Gross Domestic Product at market prices that were prevailing during the time that it was calculated. For example, GDP estimate for 1974 was calculated using the value of money as at 1974.

The growth rate for year t was calculated as:

$$\text{currentGDP} = \frac{\text{Current GDP}_t - \text{Current GDP}_{t-1}}{\text{Current GDP}_{t-1}}$$

M1 money supply

M1 money supply was obtained as is from the literature; it is the most liquid form of money in the economy and it consists of currency outside banks, private demand deposits and 7 day notice time deposits.

The growth rate for year t was calculated as:

$$M1 = \frac{M1\ Money_t - M1\ Money_{t-1}}{M1\ Money_{t-1}}$$

3.3.2 Multiple Indicators Multiple Causes (MIMIC) model

The model in use for this study will be conducted in a manner that is as simple as possible, uses all variables in their stationary form, follows statistical tests and analyzes variance inherent in the best model.

$$\begin{aligned} \mathbf{Y}_i &= f(\eta_i) \\ \eta_i &= f(\mathbf{X}_i) \end{aligned} \tag{7}$$

$$\begin{aligned} Y_1 &= f(\text{GDPpc}) && \sim I(0) \\ Y_2 &= f(\text{currentGDP}) && \sim I(0) \\ Y_3 &= f(M1) && \sim I(0) \end{aligned}$$

$$\mathbf{Y}_1 = \begin{pmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,47} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,47} \\ Y_{3,1} & Y_{3,2} & \dots & Y_{3,47} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{pmatrix}$$

$$\mathbf{Y}_2 = \begin{pmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,37} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,37} \\ Y_{3,1} & Y_{3,2} & \dots & Y_{3,37} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{pmatrix}$$

$$\mathbf{Y}_3 = \begin{pmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,35} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,35} \\ Y_{3,1} & Y_{3,2} & \dots & Y_{3,35} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{pmatrix}$$

$$\eta_i = f(nIS) \sim I(0) \quad \text{or} \quad \eta_i = f(\text{any } \underline{\eta}_j \text{ where } j = 1, 2, \dots, n) \sim I(0)$$

$$\eta_1 = (\eta_1 \quad \eta_2 \quad \dots \quad \eta_{47})$$

$$\eta_2 = \begin{pmatrix} \eta_{1,1} & \eta_{1,2} & \dots & \eta_{1,37} \\ \eta_{2,1} & \eta_{2,2} & \dots & \eta_{2,37} \end{pmatrix} = \begin{pmatrix} \underline{\eta}_1 \\ \underline{\eta}_2 \end{pmatrix}$$

$$\eta_3 = \begin{pmatrix} \eta_{1,1} & \eta_{1,2} & \dots & \eta_{1,35} \\ \eta_{2,1} & \eta_{2,2} & \dots & \eta_{2,35} \\ \vdots & \vdots & \vdots & \vdots \\ \eta_{6,1} & \eta_{6,2} & \dots & \eta_{6,35} \end{pmatrix} = \begin{pmatrix} \underline{\eta}_3 \\ \underline{\eta}_4 \\ \underline{\eta}_5 \\ \underline{\eta}_6 \\ \underline{\eta}_7 \\ \underline{\eta}_8 \end{pmatrix}$$

$$X_1 = f(ITburden) \sim I(0)$$

$$X_2 = f(GovSize) \sim I(0)$$

$$X_3 = f(ProxyUnem) \sim I(0)$$

$$X_4 = f(TimeDepIR) \sim I(0)$$

$$\mathbf{X}_1 = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,47} \\ X_{2,1} & X_{2,2} & \dots & X_{2,47} \\ \vdots & \vdots & & \vdots \\ X_{4,1} & X_{4,2} & \dots & X_{4,47} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_5 \end{pmatrix}$$

$$\mathbf{X}_2 = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,37} \\ X_{2,1} & X_{2,2} & \dots & X_{2,37} \\ \vdots & \vdots & & \vdots \\ X_{4,1} & X_{4,2} & \dots & X_{4,37} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_5 \end{pmatrix}$$

$$\mathbf{X}_3 = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,35} \\ X_{2,1} & X_{2,2} & \dots & X_{2,35} \\ \vdots & \vdots & & \vdots \\ X_{4,1} & X_{4,2} & \dots & X_{4,35} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_5 \end{pmatrix}$$

$$\xi_1 = \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} & \dots & \varepsilon_{1,47} \\ \varepsilon_{2,1} & \varepsilon_{2,2} & \dots & \varepsilon_{2,47} \\ \varepsilon_{3,1} & \varepsilon_{3,2} & \dots & \varepsilon_{3,47} \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

$$\xi_2 = \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} & \dots & \varepsilon_{1,37} \\ \varepsilon_{2,1} & \varepsilon_{2,2} & \dots & \varepsilon_{2,37} \\ \varepsilon_{3,1} & \varepsilon_{3,2} & \dots & \varepsilon_{3,37} \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

$$\xi_3 = \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} & \dots & \varepsilon_{1,35} \\ \varepsilon_{2,1} & \varepsilon_{2,2} & \dots & \varepsilon_{2,35} \\ \varepsilon_{3,1} & \varepsilon_{3,2} & \dots & \varepsilon_{3,35} \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

$$\mathbf{U}_1 = (\varepsilon_1 \quad \varepsilon_2 \quad \dots \quad \varepsilon_{47})$$

$$\mathbf{U}_2 = \begin{pmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} & \dots & \varepsilon_{1,37} \\ \varepsilon_{2,1} & \varepsilon_{2,2} & \dots & \varepsilon_{2,37} \end{pmatrix}$$

$$\mathbf{U}_3 = \begin{pmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,35} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,35} \\ \vdots & \vdots & \vdots & \vdots \\ \epsilon_{6,1} & \epsilon_{6,2} & \dots & \epsilon_{6,35} \end{pmatrix}$$

$$\lambda_1 = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}$$

$$\lambda_2 = \begin{pmatrix} \lambda_{1,1} & \lambda_{1,2} \\ \lambda_{2,1} & \lambda_{2,2} \\ \lambda_{3,1} & \lambda_{3,2} \end{pmatrix}$$

$$\lambda_3 = \begin{pmatrix} \lambda_{1,1} & \lambda_{1,2} & \dots & \lambda_{1,6} \\ \lambda_{2,1} & \lambda_{2,2} & \dots & \lambda_{2,6} \\ \lambda_{3,1} & \lambda_{3,2} & \dots & \lambda_{3,6} \end{pmatrix}$$

$$\beta = (\beta_1 \quad \beta_2 \quad \dots \quad \beta_4)$$

Assuming the informal sector is homogeneous and $T > 30$, the MIMIC model in the multivariate case becomes

$$\begin{aligned} \mathbf{Y}_1 &= \lambda_1 \eta_1 + \xi_1 \\ \eta_1 &= \beta \mathbf{X}_1 + \mathbf{U}_1 \end{aligned} \tag{8}$$

Assuming the informal sector is heterogeneous from a location standpoint and $T > 30$, the MIMIC model in the multivariate case becomes

$$\begin{aligned} \mathbf{Y}_2 &= \lambda_2 \eta_2 + \xi_2 \\ \eta_2 &= \beta \mathbf{X}_2 + \mathbf{U}_2 \end{aligned} \tag{9}$$

Assuming the informal sector is heterogeneous from an industry standpoint and $T > 30$, the MIMIC model in the multivariate case becomes

$$\begin{aligned} \mathbf{Y}_3 &= \lambda_3 \eta_3 + \xi_3 \\ \eta_3 &= \beta \mathbf{X}_3 + \mathbf{U}_3 \end{aligned} \tag{10}$$

Model Assumptions

Regarding the error component of the model, the two error sub components are random, homoskedastic and independent from one another i.e the error in the first equation is not dependent on the error in the second equation.

$$\begin{aligned} E(\xi) &= 0 \sim N(0, \Psi) & \text{and} & & E(\mathbf{U}) &= 0 \sim N(0, \sigma^2 \mathbf{I}) \\ \xi &\sim I(0) & \text{and} & & \mathbf{U} &\sim I(0) \\ E(\xi | \mathbf{U}) &= 0 \end{aligned}$$

Giles (1999) expressed the MIMIC model as a type of mixed model, $\mathbf{y} = \Pi \mathbf{x} + \mathbf{z}$; in this multivariate context, this would be expressed as $\mathbf{Y} = \Pi \mathbf{X} + \mathbf{Z}$ where $\Pi = \lambda \beta'$. This would imply that the covariance matrix for the error is a mixture of the estimates from the first equation as well as variance attributed to both error components i.e. $\Omega = \lambda \lambda' + \Psi$ (Krishnakumar and Nagar, 2008).

3.3.3 Data Validation

Upon inspection of the spreadsheets, the variables changed over the years i.e. variable X for year t had more than one estimate. To deal with this, all the possible estimates of the results for each variable, each year were collected and a measure of central tendency, their average, could be used as the final estimate for each variable, each year.

Regarding the issue of missing values, those values were handled differently depending on the dataset. In the 1972-2019 panel dataset, there was only one missing value and it was in 1972; therefore the 1972 observation was removed and the data under study was for the years 1973-2019. In the 1983-2019 panel dataset, there were 13 missing values for the years 1983-1997 and they were found to all belong to the *Other Industries* category. Therefore, a measure of central tendency, the median of all values in that particular industry could be used to fill the missing values, ensuring a balanced panel dataset. The 1985-2019 panel dataset was found to have no missing values and was therefore left as is.

Stationarity in the variables and their various transformations could be checked using

the Augmented Dickey-Fuller (ADF) Unit Root test (Wooldridge, 2013) shown below:

$$\text{Model: } x_t = \mu + \theta_1 x_{t-1} + \theta_2 x_{t-2} + \theta_3 x_{t-3} + \varepsilon_t$$

$$H_0 : \theta = 1 \text{ i.e. } X \not\sim I(0)$$

$$H_1 : \theta < 1 \text{ i.e. } X \sim I(0)$$

$\alpha = 0.05$ level of statistical significance

$$\text{Test statistic: } t_{(\theta=1)} = \frac{\hat{\theta} - 1}{s.e.(\hat{\theta})}$$

Normality in the individual variables could be checked through the Shapiro and Wilk (1965) test for normality, which is as follows:

$$H_0 : (x_1, x_2, \dots, x_t) \sim N(\mu, \sigma^2)$$

$$H_1 : (x_1, x_2, \dots, x_t) \not\sim N(\mu, \sigma^2)$$

$\alpha = 0.05$ level of statistical significance

$$\text{Test statistic: } W = \frac{(\sum_{i=1}^t a_i x_{(i)})^2}{\sum_{i=1}^t (x_i - \bar{x})^2}$$

Each set of panel datasets that formed a matrix containing individual variables following the normal distribution could be tested for multivariate normality. This would involve using Mardia's test that simultaneously checks for skewness and kurtosis in the matrix (Rencher, 2003). The test when applied on the cause variables would be as follows:

$$H_0 : \mathbf{X}' = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_p \end{pmatrix} \sim N_p(\mu, \Sigma)$$

$$H_1 : \mathbf{X}' = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_p \end{pmatrix} \not\sim N_p(\mu, \Sigma)$$

$\alpha = 0.05$ level of statistical significance

$$\text{Test statistics: } \beta_{1,p} = E \left[(\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu) \right]^3$$

$$\beta_{2,p} = E \left[(\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu) \right]^2$$

$$\hat{\Sigma} = \frac{1}{T} \sum_{i=1}^T (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

$$g_{i,j} = (\mathbf{x}_i - \mu)' \hat{\Sigma}^{-1} (\mathbf{x}_j - \mu)$$

$$\hat{\beta}_{1,p} = \frac{1}{T^2} \sum_{i=1}^T \sum_{j=1}^T g_{i,j}^3 \quad \text{and} \quad \hat{\beta}_{2,p} = \frac{1}{T} \sum_{i=1}^T g_{i,i}^2$$

$$z_1 = \frac{(p+1)(T+1)(T+3)}{6[(T+1)(p+1)-6]} \hat{\beta}_{1,p} \quad \sim \quad \chi^2 \left(\frac{1}{6} p(p+1)(p+2) \right)$$

$$z_2 = \frac{\hat{\beta}_{2,p} - p(p+2)(T+p+1)}{\sqrt{\frac{8p(p+2)}{(T-1)}}} \quad \sim \quad N(0,1)$$

Finally, ideally there should be little to no multicollinearity within the cause variables; therefore the variance inflation factor for each variable should ideally be less than 3.3 (Posey et al., 2015). Fox (2016) gave it as

$$VIF = \frac{1}{1 - R_{\text{variable}_j}^2}$$

when variable j is regressed on the other variables in a particular group.

3.4 Parameter Estimation

Parameter estimation to obtain solutions to a MIMIC model can be approached in two ways: using maximum likelihood estimation or least-square estimation.

3.4.1 Maximum Likelihood

To have valid estimates, the data of \mathbf{Y} , \mathbf{X} and $\boldsymbol{\eta}$ must be multivariate normal, with no presence of skewness or kurtosis. Rencher (2003) in a discussion of multivariate normal data describes the multivariate normal density function as

$$g(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\boldsymbol{\mu})} \quad \sim \quad N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

p = number of variables

$$\mathbf{y}' = (y_1 \quad y_2 \quad \dots \quad y_p)$$

$\boldsymbol{\Sigma}$ = $p \times p$ covariance matrix for \mathbf{y}

$$\boldsymbol{\mu}' = (\mu_1 \quad \mu_2 \quad \dots \quad \mu_p)$$

$$(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \text{Mahalanobis Distance}$$

Expressing \mathbf{Y} , $\boldsymbol{\eta}$ and \mathbf{X} in this context becomes

$$\mathbf{Y} = \begin{pmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,T} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,T} \\ Y_{3,1} & Y_{3,2} & \dots & Y_{3,T} \end{pmatrix} = (\mathbf{y}_1 \quad \mathbf{y}_2 \quad \dots \quad \mathbf{y}_T) \quad \sim \quad N_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where \mathbf{y}_1 , \mathbf{y}_2 , and \mathbf{y}_T are observation vectors

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_{1,1} & \eta_{1,2} & \dots & \eta_{1,T} \\ \vdots & \vdots & & \vdots \\ \eta_{p,1} & \eta_{p,2} & \dots & \eta_{p,T} \end{pmatrix} = (\eta_1 \quad \eta_2 \quad \dots \quad \eta_T) \quad \sim \quad N_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where η_1 , η_2 , ..., η_T are observation vectors

$$\mathbf{X} = \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,T} \\ \vdots & \vdots & & \vdots \\ X_{4,1} & X_{4,2} & \dots & X_{4,T} \end{pmatrix} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_T) \quad \sim \quad N_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where \mathbf{x}_1 , \mathbf{x}_2 , ..., \mathbf{x}_T are observation vectors

The Maximum Likelihood estimation approach to solve the simultaneous equations (Rencher and Schaalje, 2008) would be obtained as follows:

Maximize $L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \mathbf{I})$

$$(\eta_1 \quad \eta_2 \quad \dots \quad \eta_T) \quad \sim \quad N_T(\boldsymbol{\beta} \mathbf{X}, \boldsymbol{\sigma}^2 \mathbf{I})$$

$$\begin{aligned}
L(\beta, \sigma^2 \mathbf{I}) &= f(\eta; \beta, \sigma^2) \\
&= \frac{1}{(2\pi)^{T \times \frac{1}{2}} |\sigma^2|^{\frac{1}{2}}} e^{-(\eta - \beta \mathbf{X})' (\sigma^2)^{-1} (\eta - \beta \mathbf{X}) \frac{1}{2}} \\
&= \prod_{t=1}^T f(\eta_t; \beta \mathbf{x}_t, \sigma^2) \\
&= \prod_{t=1}^T \frac{1}{(2\pi)^{\frac{T}{2}} |\sigma^2|^{\frac{1}{2}}} e^{-(\eta_t - \beta \mathbf{x}_t)' (\sigma^2)^{-1} (\eta_t - \beta \mathbf{x}_t) \frac{1}{2}}
\end{aligned}$$

$$\therefore \text{ If } L(\beta, \sigma^2 \mathbf{I}) = \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} e^{-\frac{(\eta - \beta \mathbf{X})' (\eta - \beta \mathbf{X})}{2\sigma^2}}$$

$$\begin{aligned}
\ln[L(\beta, \sigma^2 \mathbf{I})] &= -\frac{T}{2} \ln[2\pi] - \frac{T}{2} \ln[\sigma^2] - \frac{1}{2\sigma^2} (\eta - \beta \mathbf{X})' (\eta - \beta \mathbf{X}) \\
&= -\frac{T}{2} \ln[2\pi] - \frac{T}{2} \ln[\sigma^2] - \frac{1}{2\sigma^2} [\eta' \eta - 2\mathbf{X}' \beta' \eta + (\beta \mathbf{X})' (\beta \mathbf{X})]
\end{aligned}$$

$$\frac{\ln[L(\beta, \sigma^2 \mathbf{I})]}{\delta \beta} = 1(-2\mathbf{X}' \beta^{1-1=0} \eta) + 2(\beta^{2-1=1} \mathbf{X}^2) = 0$$

$$\frac{\ln[L(\beta, \sigma^2 \mathbf{I})]}{\delta \beta} = -2\eta \mathbf{X}' + 2\beta (\mathbf{X} \mathbf{X}') = 0$$

$$+2\eta \mathbf{X}' - 2\eta \mathbf{X}' + 2\beta (\mathbf{X} \mathbf{X}') = 0 + 2\eta \mathbf{X}'$$

$$\frac{2(\beta \mathbf{X} \mathbf{X}')}{2} = \frac{2\eta \mathbf{X}'}{2}$$

$$(\mathbf{X} \mathbf{X}')^{-1} \times \beta \mathbf{X} \mathbf{X}' = \eta \mathbf{X}' \times (\mathbf{X} \mathbf{X}')^{-1}$$

$$\beta = \eta \mathbf{X}' \times (\mathbf{X} \mathbf{X}')^{-1}$$

Therefore the β estimates are obtained as shown:

$$\hat{\beta} = \eta \mathbf{X}' (\mathbf{X} \mathbf{X}')^{-1} \quad (11)$$

Differentiating on the basis of σ^2 results in

$$\ln[L(\beta, \sigma^2 \mathbf{I})] = -\frac{T}{2} \ln[2\pi] - \frac{T}{2} \ln[\sigma^2] - \frac{1}{2\sigma^2} (\eta - \beta \mathbf{X})' (\eta - \beta \mathbf{X})$$

$$\frac{\ln[L(\beta, \sigma^2 \mathbf{I})]}{\delta \sigma^2} = -\left(\frac{T}{2} \times \frac{1}{\sigma^2}\right) - \left(\frac{1}{2\sigma^2} \times -\frac{1}{\sigma^2} (\eta - \beta \mathbf{X})' (\eta - \beta \mathbf{X})\right) = 0$$

$$\begin{aligned} \frac{\ln[L(\beta, \sigma^2 \mathbf{I})]}{\delta \sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\eta - \beta \mathbf{X})'(\eta - \beta \mathbf{X}) = 0 \\ +\frac{T}{2\sigma^2} - \frac{T}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\eta - \beta \mathbf{X})'(\eta - \beta \mathbf{X}) &= 0 + \frac{T}{2\sigma^2} \\ \frac{2(\sigma^2)^2}{1} \times \frac{1}{2(\sigma^2)^2}(\eta - \beta \mathbf{X})'(\eta - \beta \mathbf{X}) &= \frac{T}{2\sigma^2} \times \frac{2(\sigma^2)^2}{1} \\ \frac{1}{T} \times (\eta - \beta \mathbf{X})'(\eta - \beta \mathbf{X}) &= T\sigma^2 \times \frac{1}{T} \end{aligned}$$

Therefore, the estimated variance in the measurement equation becomes

$$\hat{\sigma}^2 = \frac{1}{T}(\eta - \beta \mathbf{X})'(\eta - \beta \mathbf{X}) \quad (12)$$

To account for the bias that can arise due to the data in use being collected from a sample (Cochran, 1977), $\hat{\sigma}^2$ is replaced by $\hat{\mathbf{S}}_1^2$,

$$\hat{\mathbf{S}}_1^2 = \frac{1}{T-2}(\eta - \beta \mathbf{X})'(\eta - \beta \mathbf{X}) \quad (13)$$

The same process is repeated for the second equation:

$$\hat{\lambda} = \mathbf{Y}\eta' (\eta\eta')^{-1} \quad (14)$$

$$\hat{\Psi} = \frac{1}{T}(\mathbf{Y} - \lambda\eta)'(\mathbf{Y} - \lambda\eta) \quad (15)$$

To account for biases like before, $\hat{\Psi}$ is replaced by $\hat{\mathbf{S}}_2^2$,

$$\hat{\mathbf{S}}_2^2 = \frac{1}{T-2}(\mathbf{Y} - \lambda\eta)'(\mathbf{Y} - \lambda\eta) \quad (16)$$

3.4.2 Ordinary Least Squares

An alternative method of estimation if \mathbf{Y} , η and \mathbf{X} are not multivariate normal is least squares estimation. Rencher and Schaalje (2008) discuss it as aiming to minimize the sum of squared differences between the original response variable and the predicted response

variable. The variables in this context would become

$$\mathbf{Y} = \begin{pmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,T} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,T} \\ Y_{3,1} & Y_{3,2} & \dots & Y_{3,T} \end{pmatrix}$$

$$\eta_1 = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \eta_{1,1} & \eta_{1,2} & \dots & \eta_{1,T} \\ \vdots & \vdots & & \vdots \\ \eta_{N,1} & \eta_{N,2} & \dots & \eta_{N,T} \end{pmatrix}$$

$$\lambda = \begin{pmatrix} \lambda_{1,0} & \lambda_{1,1} & \dots & \lambda_{1,(N+1)} \\ \lambda_{2,0} & \lambda_{2,1} & \dots & \lambda_{2,(N+1)} \\ \lambda_{3,0} & \lambda_{3,1} & \dots & \lambda_{3,(N+1)} \end{pmatrix}$$

$$\eta_2 = \begin{pmatrix} \eta_{1,1} & \eta_{1,2} & \dots & \eta_{1,T} \\ \vdots & \vdots & & \vdots \\ \eta_{N,1} & \eta_{N,2} & \dots & \eta_{N,T} \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{1,1} & X_{1,2} & \dots & X_{1,T} \\ \vdots & \vdots & & \vdots \\ X_{4,1} & X_{4,2} & \dots & X_{4,T} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_{1,0} & \beta_{1,1} & \dots & \beta_{1,5} \\ \vdots & \vdots & & \vdots \\ \beta_{N,0} & \beta_{N,1} & \dots & \beta_{N,5} \end{pmatrix}$$

Therefore based on the previous matrices, the errors would be minimized as follows:

$$\mathbf{Y} = \lambda \eta_1 + \xi$$

$$\mathbf{Y} = \hat{\mathbf{Y}} + \xi$$

$$\xi = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$\xi = \mathbf{Y} - \lambda \eta_1$$

$$\eta_2 = \beta \mathbf{X} + \mathbf{U}$$

$$\eta_2 = \hat{\eta}_2 + \mathbf{U}$$

$$\mathbf{U} = \eta_2 - \hat{\eta}_2$$

$$\mathbf{U} = \eta_2 - \beta \mathbf{X}$$

$$(\xi)^2 = (\mathbf{Y} - \lambda \eta_1)^2$$

$$\hat{\xi}^T \hat{\xi} = (\mathbf{Y} - \hat{\lambda} \eta_1)^T (\mathbf{Y} - \hat{\lambda} \eta_1)$$

$$\hat{\xi}^T \hat{\xi} = \mathbf{Y}'\mathbf{Y} - 2\eta_1' \hat{\lambda}' \mathbf{Y} + (\hat{\lambda} \eta_1)' (\hat{\lambda} \eta_1)$$

$$(\mathbf{U})^2 = (\eta_2 - \beta \mathbf{X})^2$$

$$\hat{\mathbf{U}}^T \hat{\mathbf{U}} = (\eta_2 - \hat{\beta} \mathbf{X})^T (\eta_2 - \hat{\beta} \mathbf{X})$$

$$\hat{\mathbf{U}}^T \hat{\mathbf{U}} = \eta_2' \eta_2 - 2\mathbf{X}' \hat{\beta}' \eta_2 + (\hat{\beta} \mathbf{X})' (\hat{\beta} \mathbf{X})$$

Minimize $\hat{\xi}^T \hat{\xi}$:

$$\frac{\delta \hat{\xi}^T \hat{\xi}}{\delta \hat{\lambda}} = 0 - 2\mathbf{Y} \eta_1' + 2\hat{\lambda} \eta_1 \eta_1' = 0$$

$$+ 2\mathbf{Y} \eta_1' - 2\mathbf{Y} \eta_1' + 2\hat{\lambda} \eta_1 \eta_1' = +2\mathbf{Y} \eta_1'$$

$$2\hat{\lambda} \eta_1 \eta_1' = 2\mathbf{Y} \eta_1'$$

$$\frac{2\hat{\lambda} \eta_1 \eta_1'}{2\eta_1 \eta_1'} = \frac{2\mathbf{Y} \eta_1'}{2\eta_1 \eta_1'}$$

$$\hat{\lambda} = \mathbf{Y} \eta_1' (\eta_1 \eta_1')^{-1}$$

The above result is the similar to that in Equation 14. $\hat{\Psi}$ would be estimated as follows:

$$\hat{\Psi} = E[\mathbf{Y} - \hat{\mathbf{Y}}]^2$$

$$\hat{\Psi} = E[\mathbf{Y} - E(\mathbf{Y})]^2$$

$$\hat{\Psi} = E[\mathbf{Y} - \hat{\lambda} \eta_1]^2$$

$$\widehat{\Psi} = (\mathbf{Y} - \widehat{\lambda}\eta_1)'(\mathbf{Y} - \widehat{\lambda}\eta_1)$$

$\widehat{\Psi}$ is typically not used due to bias. $\widehat{\mathbf{S}}_1$ is used instead and it is obtained as:

$$\begin{aligned}\widehat{\mathbf{S}}_1^2 &= \frac{1}{T - k_1 - 1} SSE \\ \widehat{\mathbf{S}}_1^2 &= \frac{1}{T - k_1 - 1} \mathbf{Y}\mathbf{Y}' - \lambda\eta_1\mathbf{Y}'\end{aligned}\quad (17)$$

When the same process is repeated to estimate β and σ^2 ,

$$\begin{aligned}\widehat{\beta} &= \eta_2\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1} \\ \widehat{\sigma}^2 &= (\eta_2 - \beta\mathbf{X})'(\eta_2 - \beta\mathbf{X}) \\ \widehat{\mathbf{S}}_2^2 &= \frac{1}{T - k_2 - 1} \eta_2\eta_2' - \beta\mathbf{X}\eta_2'\end{aligned}\quad (18)$$

3.4.3 Statistical significance and precision of estimates

Statistical significance of the estimated parameters could be determined as follows:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

$\alpha = 0.05$ level of statistical significance

$$\text{Test statistic: } \hat{t}_\beta = \frac{\widehat{\beta}_i - 0}{s.e.(\widehat{\beta}_i)} \sim t(T)$$

$$H_0 : \lambda_i = 0$$

$$H_1 : \lambda_i \neq 0$$

$\alpha = 0.05$ level of statistical significance

$$\text{Test statistic: } \hat{t}_\lambda = \frac{\widehat{\lambda}_i - 0}{s.e.(\widehat{\lambda}_i)} \sim t(T)$$

Precision of the estimated parameters could be determined as follows:

$$\alpha = 0.05 \text{ level of statistical significance}$$

$$\hat{t}_\beta = \frac{\hat{\beta}_i - 0}{s.e.(\beta_i)}$$

$$C.I_\alpha = \hat{t}_\beta \pm t_{\frac{\alpha}{2}} s.e.(\beta)$$

$$\hat{t}_\lambda = \frac{\hat{\lambda}_i - 0}{s.e.(\lambda_i)}$$

$$C.I_\alpha = \hat{t}_\lambda \pm t_{\frac{\alpha}{2}} s.e.(\lambda)$$

3.5 Goodness of fit tests

It is recommended that the test conducted in order to prevent spurious regression be the Engle-Granger Test for cointegration. The Engle-Granger Test proved that non-stationary variables integrated of order $I(1)$ could have a relationship integrated of order $I(0)$. (Gujarati and Porter, 2009; Wooldridge, 2013).

$$\begin{aligned} \text{Model: } \epsilon_t &= \mu_1 + \theta_{1,1}\epsilon_{t-1} + \theta_{1,2}\epsilon_{t-2} + \theta_{1,3}\epsilon_{t-3} + u_{1,t} \\ \epsilon_t &= \mu_2 + \theta_{2,1}\epsilon_{t-1} + \theta_{2,2}\epsilon_{t-2} + \theta_{2,3}\epsilon_{t-3} + u_{2,t} \end{aligned}$$

$$\text{Model assumption: } \text{cor}(u_{1,t}, u_{2,t}) = 0$$

$$\begin{aligned} H_0: \quad \mathbf{Y} = \lambda \boldsymbol{\eta} + \boldsymbol{\xi} &\not\sim I(0) & \text{and} & \quad \boldsymbol{\eta} = \boldsymbol{\beta} \mathbf{X} + \mathbf{U} &\not\sim I(0) & \text{i.e.} \\ &\forall \theta_{1,i} = 1 & & \text{and} & \forall \theta_{2,i} = 1 & \\ H_1: \quad \mathbf{Y} = \lambda \boldsymbol{\eta} + \boldsymbol{\xi} &\sim I(0) & \text{and} & \quad \boldsymbol{\eta} = \boldsymbol{\beta} \mathbf{X} + \mathbf{U} &\sim I(0) & \text{i.e.} \\ &\forall \theta_{1,i} < 1 & & \text{and} & \forall \theta_{2,i} < 1 & \end{aligned}$$

$$\alpha = 0.05 \text{ level of statistical significance}$$

$$\text{Test statistic: } \hat{t}_\theta = \frac{\hat{\theta} - 1}{s.e.(\theta)}$$

The coefficient of determination R^2 for each equation within the MIMIC model could give a good indicator of goodness of fit; the higher R^2 is, the more variation is explained by the particular equation in the system. It is computed as follows (Gujarati and Porter, 2009):

$$\mathbf{Y} = \lambda \boldsymbol{\eta} + \boldsymbol{\xi} \therefore$$

$$\mathbf{Y} = \widehat{\mathbf{Y}} + \widehat{\boldsymbol{\xi}}$$

$$\boldsymbol{\eta} = \beta \mathbf{X} + \mathbf{U} \therefore$$

$$\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}} + \widehat{\mathbf{U}}$$

$$SST = SSR + SSE$$

$$\sum_{i=1}^p \sum_{t=1}^T y_{it}^2 = \sum_{i=1}^p \sum_{t=1}^T \hat{y}_{it}^2 + \sum_{i=1}^p \sum_{t=1}^T \hat{\epsilon}_{it}^2$$

$$\sum_{i=1}^q \sum_{t=1}^T \eta_{it}^2 = \sum_{i=1}^q \sum_{t=1}^T \hat{\eta}_{it}^2 + \sum_{i=1}^q \sum_{t=1}^T \hat{\epsilon}_{it}^2$$

$$R^2 = \frac{SSE}{SST} \quad (19)$$

Applied to the MIMIC model, the final R^2 criterion for the model would be

$$R^2 = \frac{\sum_{i=1}^p \sum_{t=1}^T \hat{\epsilon}_{it}^2 + \sum_{i=1}^q \sum_{t=1}^T \hat{\epsilon}_{it}^2}{\sum_{i=1}^p \sum_{t=1}^T y_{it}^2 + \sum_{i=1}^q \sum_{t=1}^T \eta_{it}^2} \quad (20)$$

In addition to this measure, Akaike Information Criterion (AIC) can indicate goodness of fit; it's formula is (Gujarati and Porter, 2009):

$$AIC = e^{2k/T} \frac{SSE}{T} \quad (21)$$

Applied to the MIMIC model, the final AIC criterion for the model would be

$$\text{Let } factor = \frac{2(k_1 + k_2)}{(T \times 2)}$$

where k_1 = number of \hat{y} regressors and k_2 = number of $\hat{\eta}$ regressors

$$AIC = e^{factor} \frac{\sum_{i=1}^p \sum_{t=1}^T \hat{\epsilon}_{it}^2 + \sum_{i=1}^q \sum_{t=1}^T \hat{\epsilon}_{it}^2}{T \times 2} \quad (22)$$

The root of the mean of SSE would be used to explain variation as derived in a multivariate analysis of variance as discussed by Rencher (2003) shown below:

Table 1. Multivariate Analysis of Variance applied to the MIMIC model.

SOURCE OF VARIATION	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SUM OF SQUARES
$\mathbf{Y} = \lambda\boldsymbol{\eta} + \boldsymbol{\xi}$ $\boldsymbol{\eta} = \boldsymbol{\beta}\mathbf{X} + \mathbf{U}$	$\sum_{i=1}^p \sum_{t=1}^T \hat{y}_{it}^2$ $\sum_{i=1}^q \sum_{t=1}^T \hat{\eta}_{it}^2$	$(p+q)$	$\frac{\sum_{i=1}^p \sum_{t=1}^T \hat{y}_{it}^2 + \sum_{i=1}^q \sum_{t=1}^T \hat{\eta}_{it}^2}{(p+q)}$
Combined Error	$\sum_{i=1}^p \sum_{t=1}^T \hat{\boldsymbol{\epsilon}}_{it}^2$ $\sum_{i=1}^q \sum_{t=1}^T \hat{\boldsymbol{\epsilon}}_{it}^2$	$SST - (p+q)$	$\frac{\sum_{i=1}^p \sum_{t=1}^T \hat{\boldsymbol{\epsilon}}_{it}^2 + \sum_{i=1}^q \sum_{t=1}^T \hat{\boldsymbol{\epsilon}}_{it}^2}{SST - (p+q)}$
TOTAL VARIATION	$\sum_{i=1}^p \sum_{t=1}^T y_{it}^2$ $\sum_{i=1}^q \sum_{t=1}^T \eta_{it}^2$	$SST =$ $2(T-2)$ or $2T - (k_1 + k_2) - 2$	$\frac{\sum_{i=1}^p \sum_{t=1}^T y_{it}^2 + \sum_{i=1}^q \sum_{t=1}^T \eta_{it}^2}{SST}$

p = number of \hat{y} variables and q = number of $\hat{\eta}$ variables .

Finally, the overall model would need to be tested for statistical significance (Rencher, 2003) as follows:

$$H_0 : (\boldsymbol{\lambda}, \boldsymbol{\beta}) = 0 \quad \text{vs.} \quad H_1 : (\boldsymbol{\lambda}, \boldsymbol{\beta}) \neq 0$$

$\alpha = 0.05$ level of statistical significance

$$\text{Test statistic: } F_c = \frac{\text{Mean Sum of Squares for Overall Regression}}{\text{Mean Sum of Squares for Combined Error}}$$

$$F_c \sim F_{\alpha/2} \left[(p+q), [2(T-2) \vee 2T - (k_1 + k_2) - 2] - (p+q) \right]$$

4 Results

4.1 Data

Missing values were handled as follows. Firstly, the 1972 observation was removed from the generalized panel, resulting in a 1973-2019 time range. Secondly, 13 missing values were found in *Other_nIS*; therefore the median of this group was imputed to ensure a balanced dataset in order to conduct the poolability test discussed later on in this chapter. 3 panel datasets were formed as follows:

Table 2. Panel Datasets Under Analysis.

DATASET	VARIABLE & DESCRIPTION	TIME RANGE
Generalized	<i>ITburden</i> Annual growth in indirect taxes	1974-2019 46 years
	<i>GovSize</i> Annual growth in public consumption	
	<i>ProxyUnem</i> Annual growth in proxy unemployment rate	
	<i>TimeDepIR</i> Annual growth in average time deposit interest rate	
	<i>nIS</i> Annual growth in estimated total number of informal sector workers	
	<i>M1</i> Annual growth in M1 money supply	
	<i>GDPpc</i> Annual growth in GDP per capita at constant prices	
	<i>currentGDP</i> Annual growth in GDP at current prices	
Location	<i>Urban_nIS</i> Annual growth in estimated total number of informal sector workers found in urban areas	1984-2019

	<i>Rural_nIS</i>	Annual growth in estimated total number of informal sector workers found in rural areas	36 years
Industry	<i>Manufacturing_nIS</i>	Annual growth in estimated total number of informal sector workers working in Manufacturing	1986-2019 34 years
	<i>Construction_nIS</i>	Annual growth in estimated total number of informal sector workers working in Construction	
	<i>TradeHospitality_nIS</i>	Annual growth in estimated total number of informal sector workers working in Trade, Hotels & Restaurants	
	<i>TransComms_nIS</i>	Annual growth in estimated total number of informal sector workers working in Transport & Communications	
	<i>CSP_nIS</i>	Annual growth in estimated total number of informal sector workers working in Community, Social & Personal Services	

4.1.1 Descriptive statistics

Data visualization was conducted using Power BI software. Each time series, with corresponding summary statistics was visualized as follows.

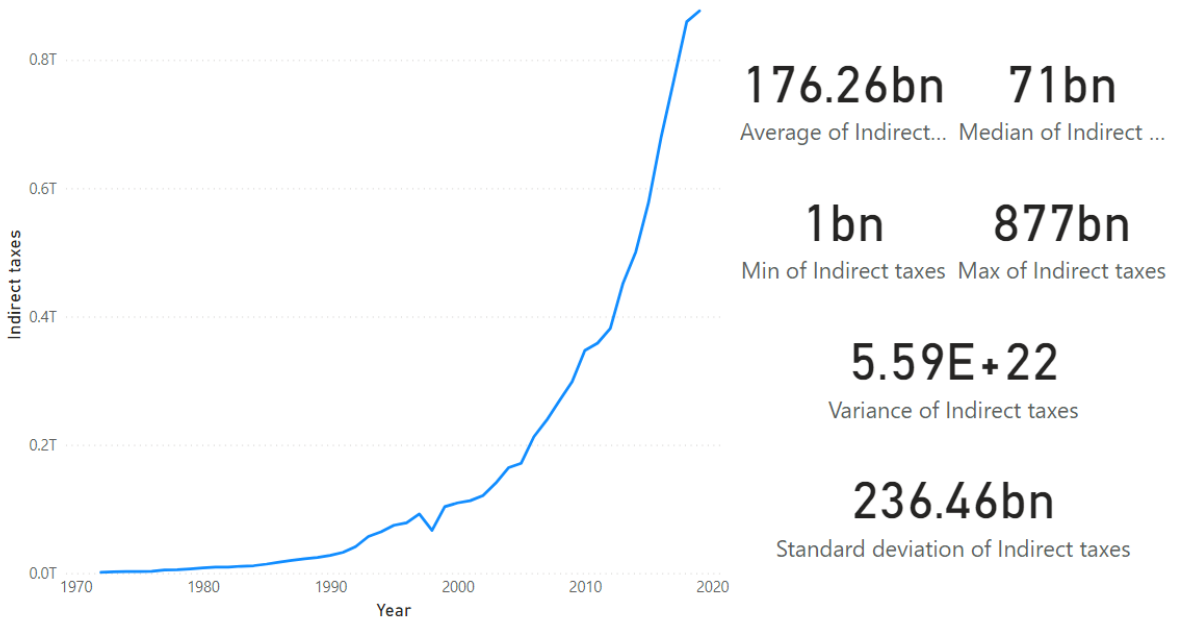


Figure 3. Total Indirect Tax over Time: Summary Statistics

Indirect tax appears to have an upward trend over time, indirect taxes being Ksh. 176B during an average year. Annual growth rates over time were:

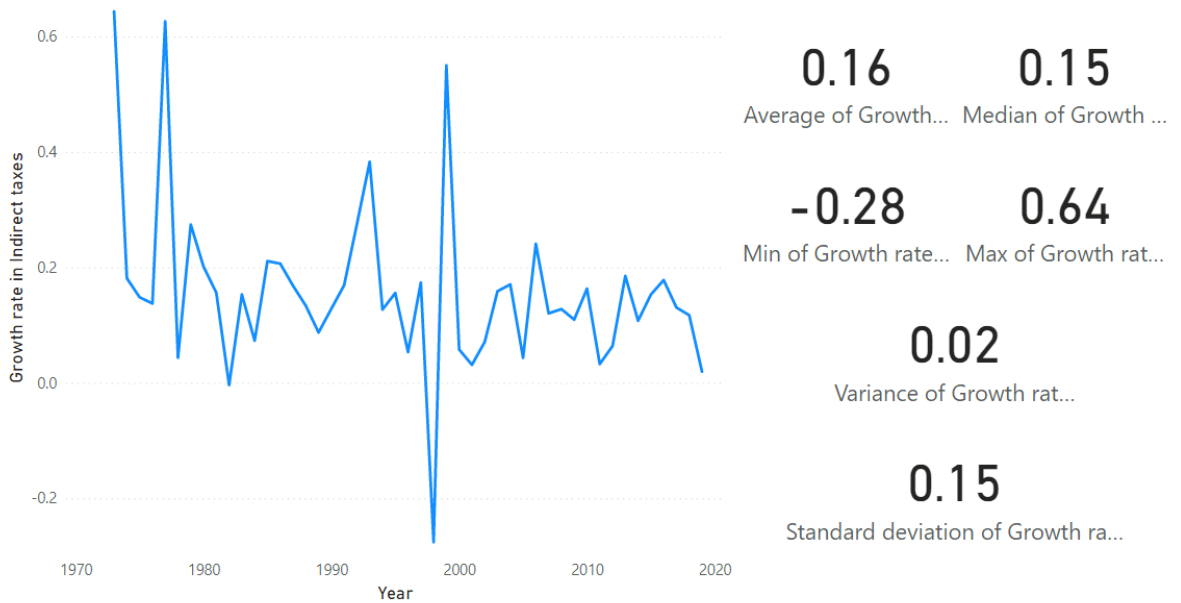


Figure 4. Growth in Total Indirect Tax over Time: Summary Statistics

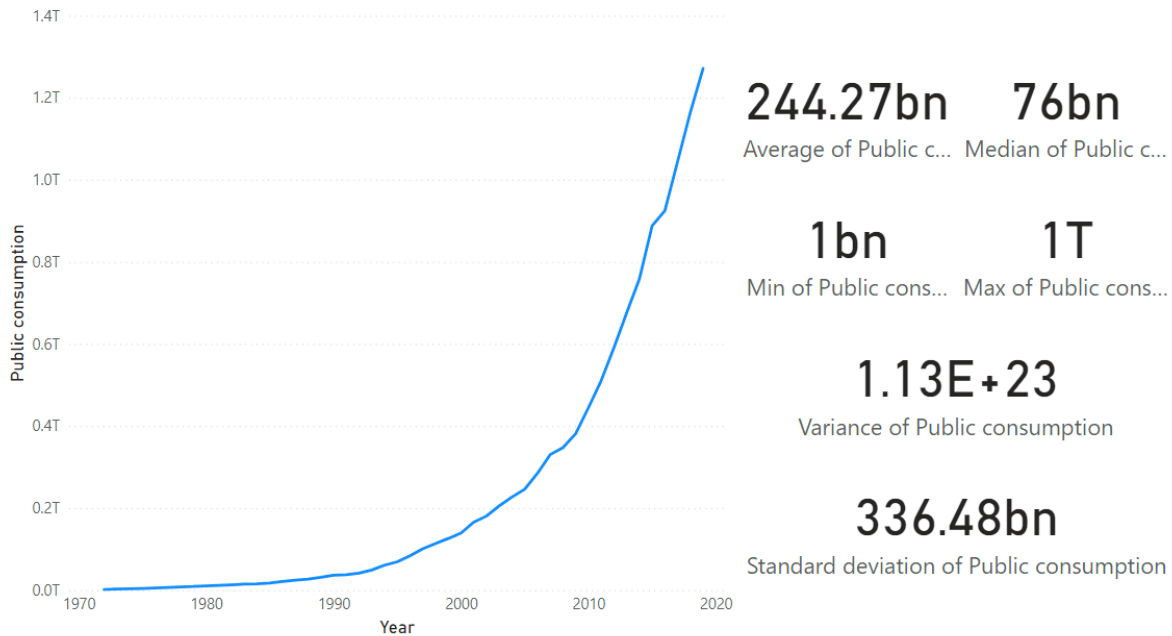


Figure 5. Total Public Consumption over Time: Summary Statistics

Public Consumption appears to have an upward trend over time, with public consumption being Ksh. 244B during an average year. Annual growth rates over time were:

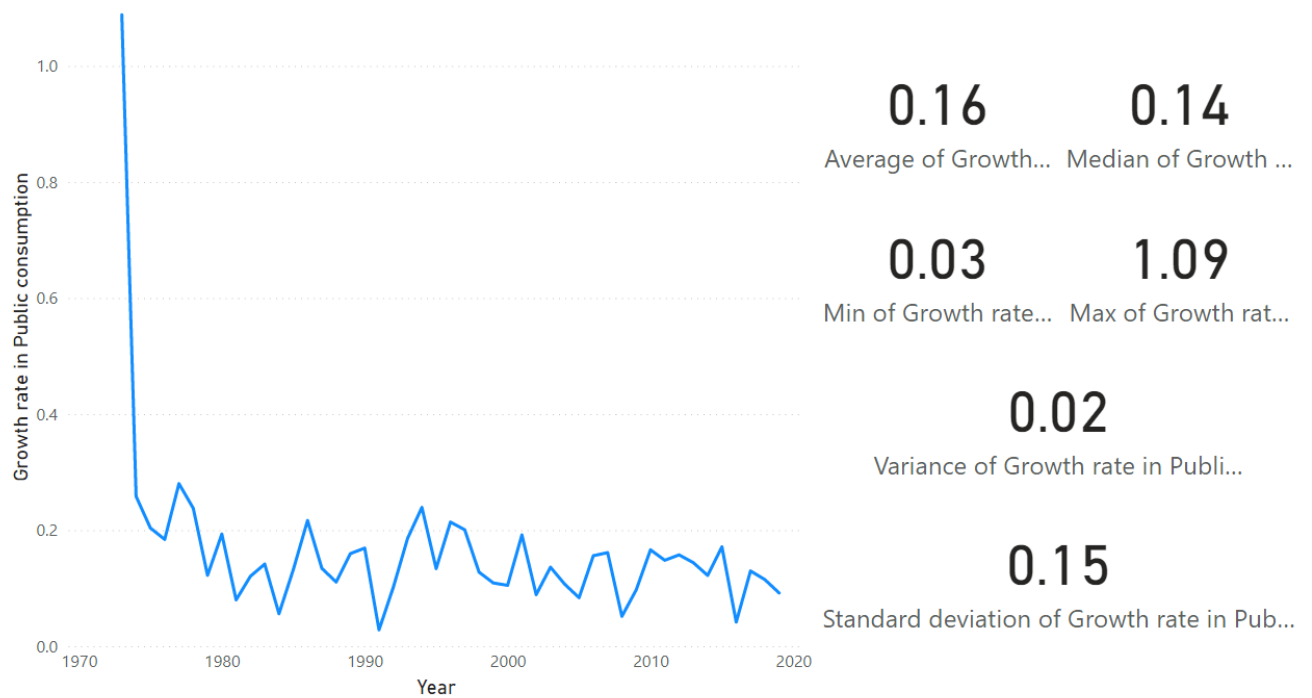


Figure 6. Growth in Total Public Consumption over Time: Summary Statistics

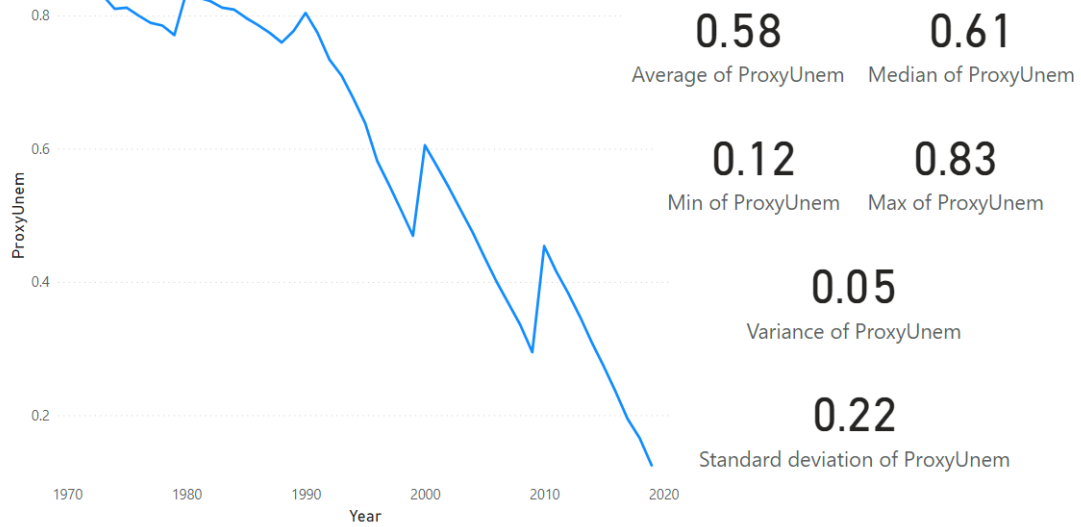


Figure 7. Proxy Unemployment Rate over Time: Summary Statistics

Proxy Unemployment Rate appears to have an downward trend over time, with spikes happening every 10 years or so. 58% of the working population are unemployed, inactive or engaged in small scale agriculture during an average year. Annual growth rates over time were:

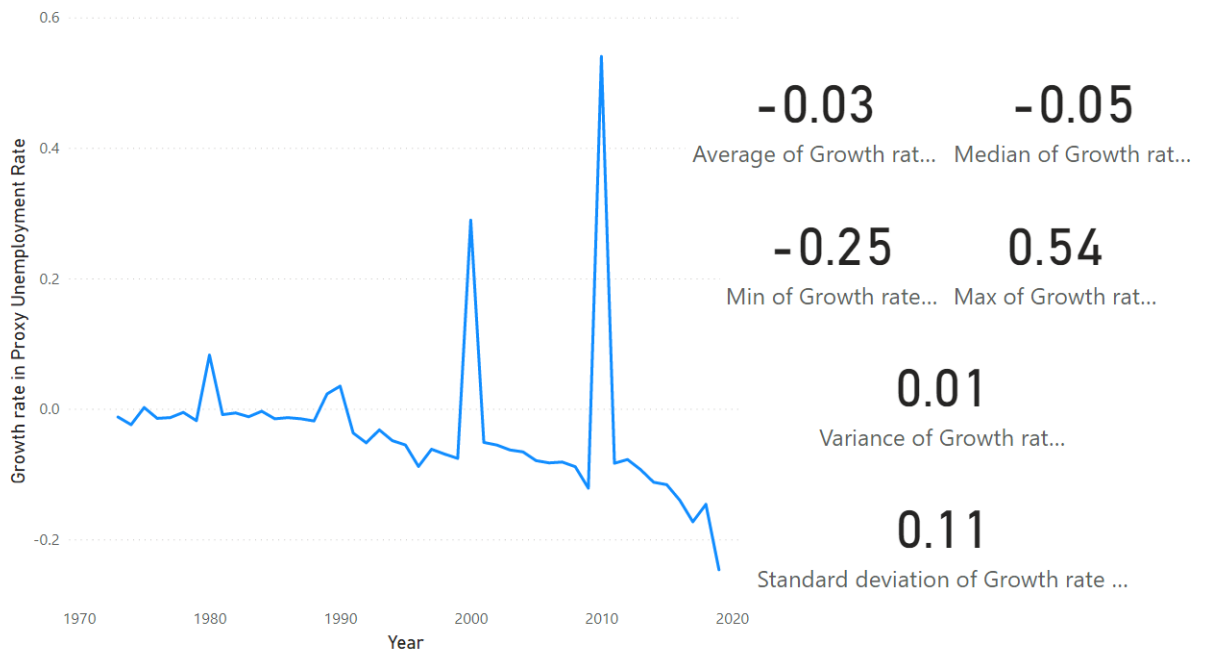


Figure 8. Growth in Proxy Unemployment Rate over Time: Summary Statistics

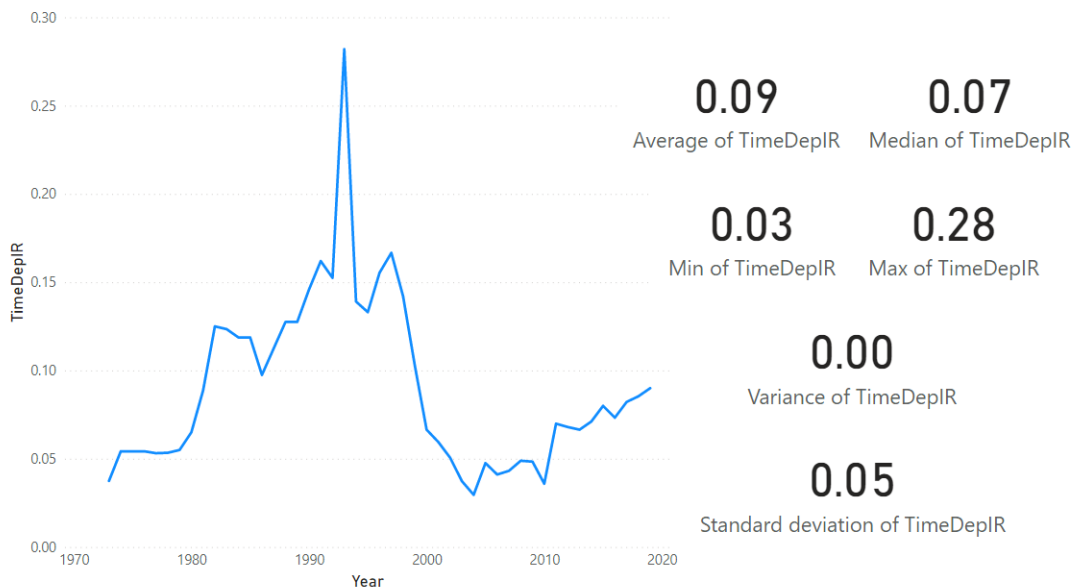


Figure 9. Average Time Deposit Interest Rate Declared by Commercial Banks over Time: Summary Statistics

Average time deposit interest rate appears to have an cyclical trend, with recent years showing an upward trend over time. It was 9% during an average year. Annual growth rates over time were:

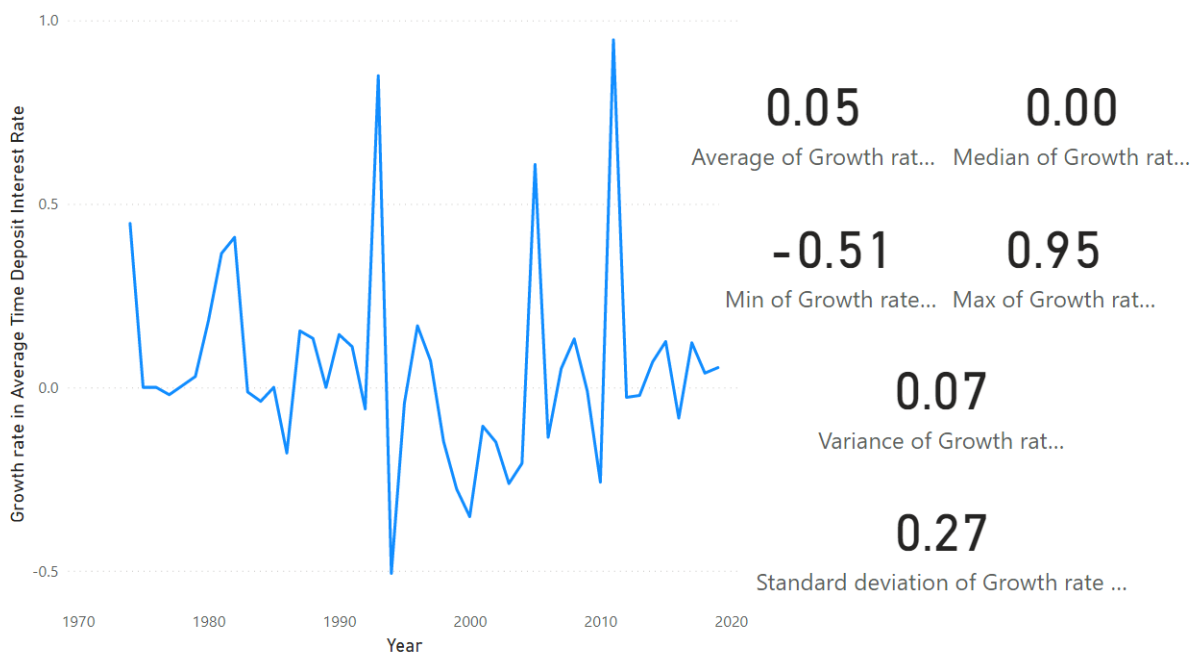


Figure 10. Growth in Average Time Deposit Interest Rate Declared by Commercial Banks over Time: Summary Statistics

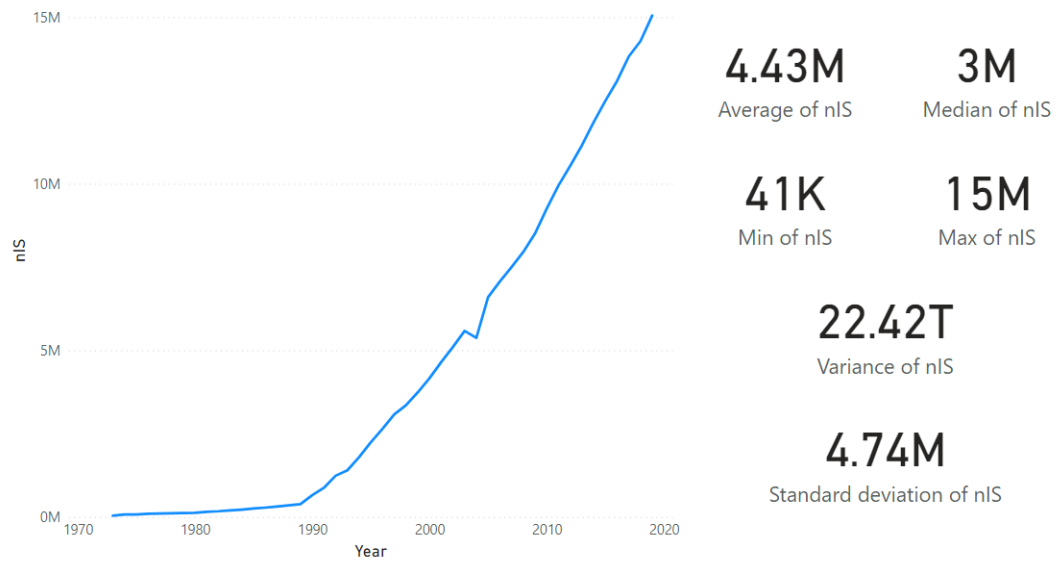


Figure 11. Number of Workers in the Informal Sector over Time: Summary Statistics

Number of workers in the informal sector appears to have an upward trend over time, with the informal sector employing 4.4 million Kenyan individuals during an average year. Annual growth rates over time were:

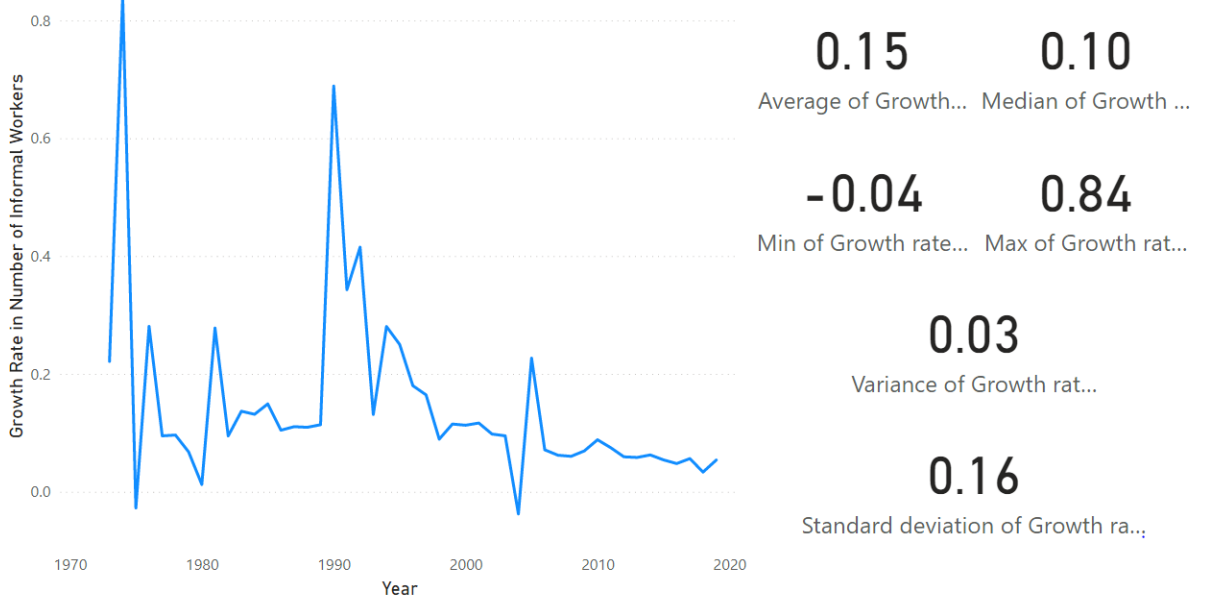


Figure 12. Growth in Number of Workers in the Informal Sector over Time: Summary Statistics

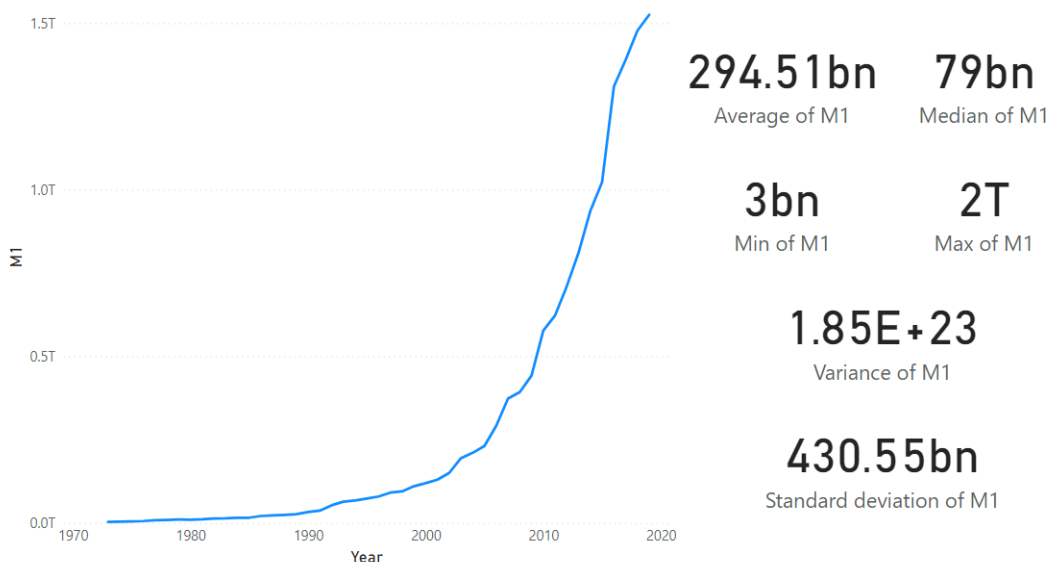


Figure 13. M1 Money Supply over Time: Summary Statistics

M1 money supply appears to have an upward trend over time. Currency outside commercial banks, demand deposits and 7 day notice time deposits were cumulatively found to be Ksh. 295 billion during an average year. Annual growth rates over time were:

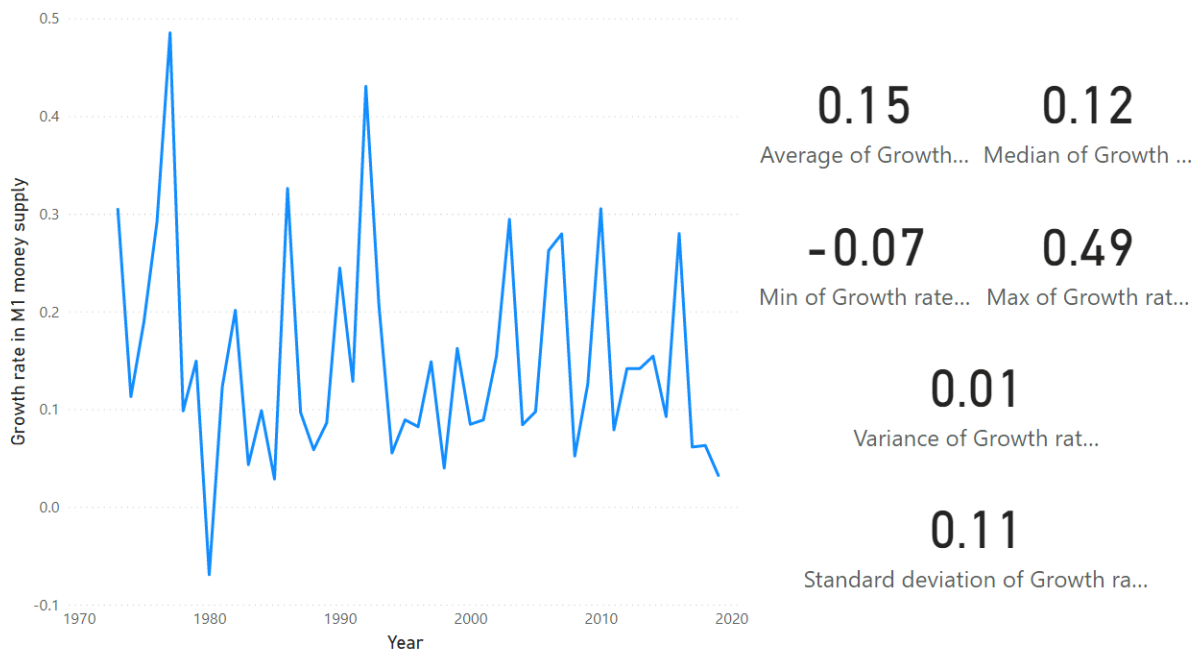


Figure 14. Growth in M1 Money Supply over Time: Summary Statistics

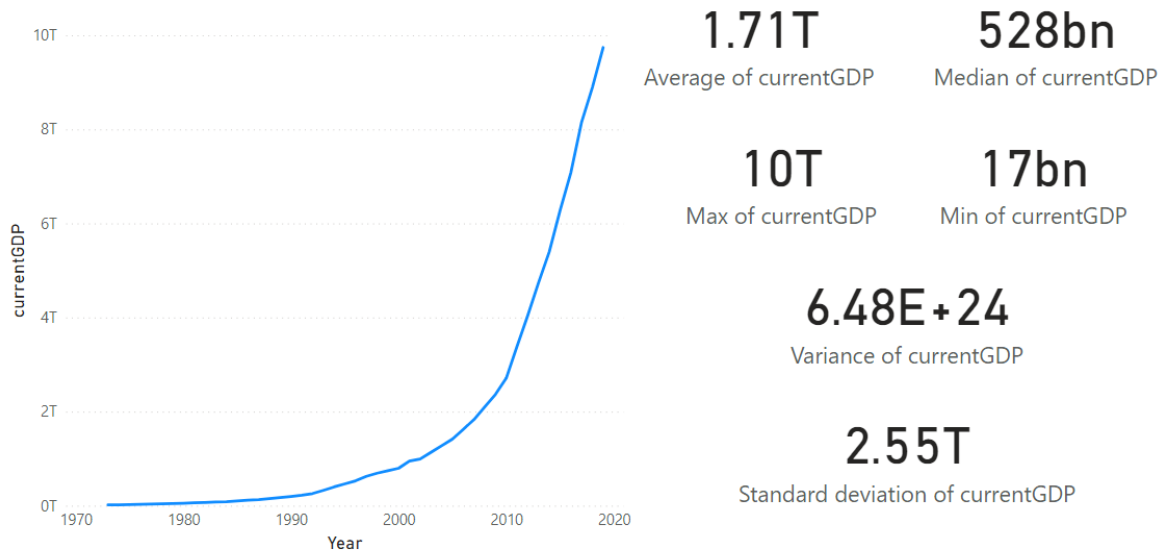


Figure 15. GDP at Current Prices over Time: Summary Statistics

GDP at current prices appears to have an upward trend over time, with Ksh. 17 trillion, current price, worth of goods and services produced during an average year. Annual growth rates over time were:

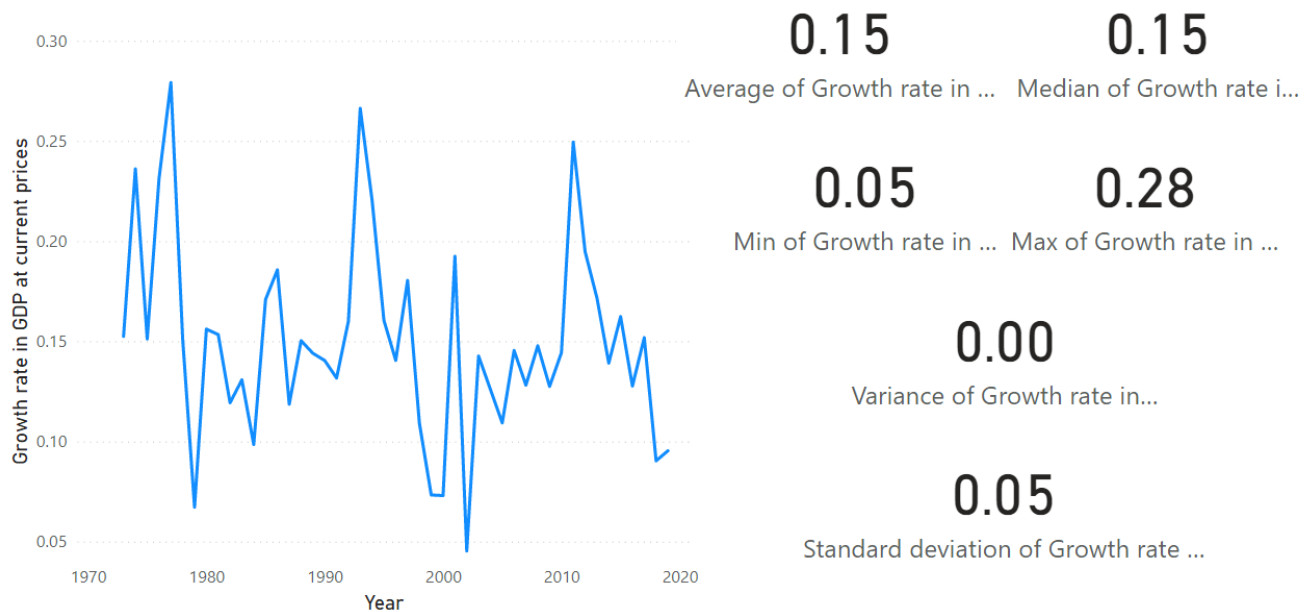


Figure 16. Growth in GDP at Current Prices over Time: Summary Statistics

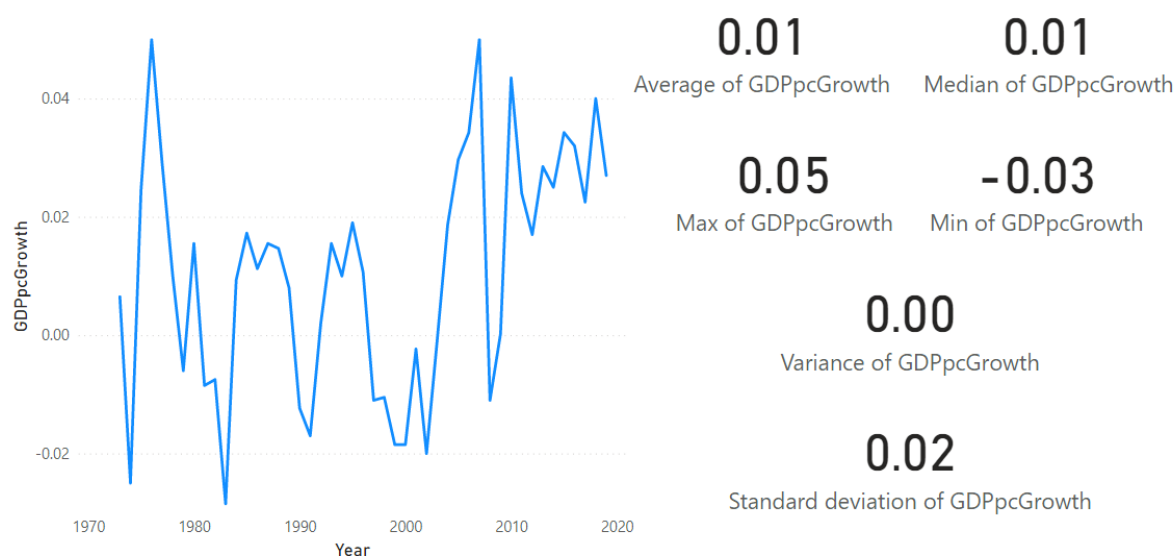


Figure 17. Growth in GDP per Capita at Constant Prices over Time: Summary Statistics

Growth in GDP per capita appears to have a slight upward trend in recent years. It was 1% during an average year.

4.2 Empirical Model

4.2.1 Heterogeneity Test Results

Tests checking for heterogeneity were conducted on the Industry and Location panel datasets. The results from the poolability test were as follows:

Table 3. Chow's (1960) poolability test results.

ASPECT BEING TESTED	F_{cal}	F_{tab}	OUTCOME
Location	45.4316	5.2406	Reject H_0 ; there are statistically significant differences between locations
Industry	58.6982	2.6279	Reject H_0 ; there are statistically significant differences between industries

Based on the results, it was worthwhile to use these datasets in place of *nIS*.

Heterogeneity based on Location

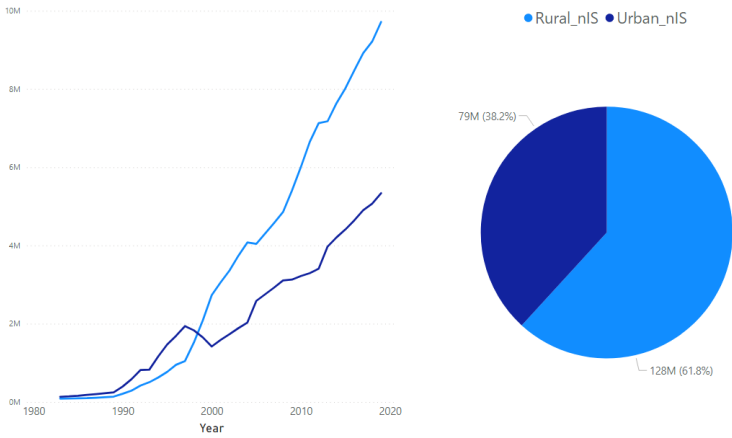


Figure 18. Comparison Between Informal Sector Workers Based on Their Location

As shown in Figure 18, the informal sector initially started quite small but has greatly grown over the years. Urban areas had the larger share of informal sector workers in the beginning, but between 1998 and 1999 a shift occurred where the number of informal sector workers in rural areas were greater than those in urban areas for the first time and the gap has only continued to widen since then. Cumulatively, most workers in the informal sector have been conducting their activities in rural areas as indicated by the pie chart above.

Summary statistics for workers in the two locations is as shown below:

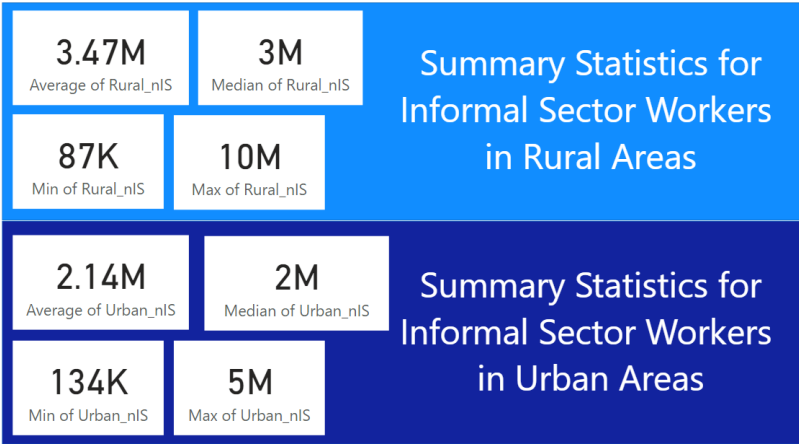


Figure 19. Summary Statistics for Informal Sector Workers Based on Location

Heterogeneity Based on Location

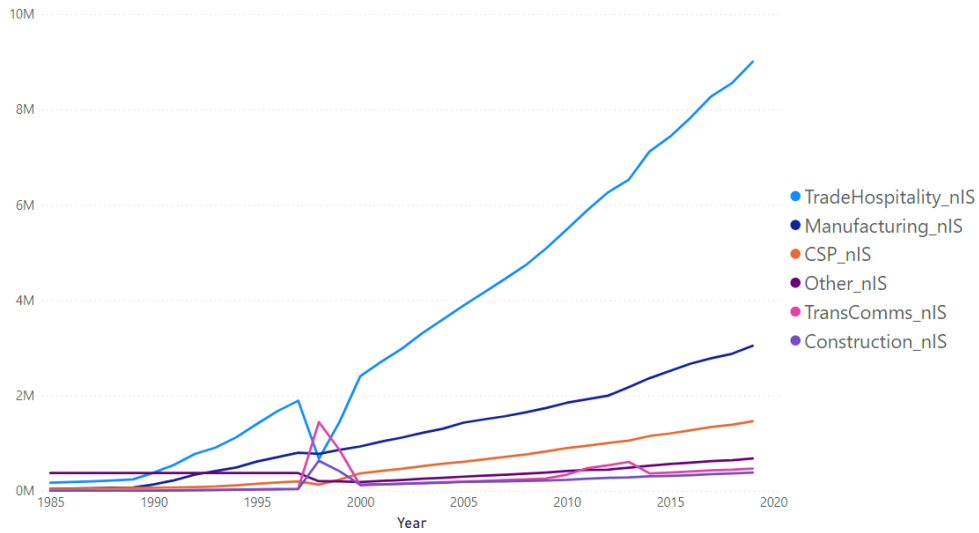


Figure 20. Comparison Between Informal Sector Workers Based on Their Industry over Time

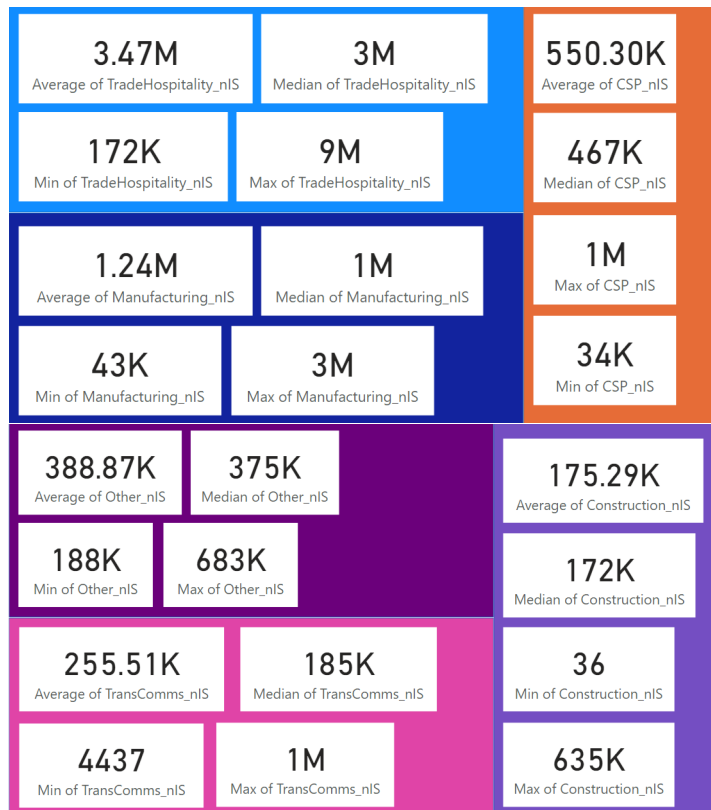


Figure 21. Summary Statistics for Informal Sector Workers Based on Industry

Figure 20 indicates the various industries operating in the informal sector over time. The industry that has experienced the most growth over time is Trade, Hotels and Restaurants, followed by Manufacturing at a distant second. Community, Social and Personal (CSP) Services is the third largest industry. Trade, Hotels and Restaurants appears to be a high growth industry, therefore, it would be noteworthy to invest its relationship to GDP per Capita and Growth in GDP per Capita. The summary statistics characterizing each industry are as visualized in Figure 21.

Data Validation Results

Data analysis and parameter estimation was conducted using R software (R Core Team, 2021). Due to the fact that *Other_nIS* had a sample size $T < 30$, it was not tested for normality or stationarity. Stationarity was checked for using *tseries* (Trapletti and Hornik, 2020) package at $\alpha = 0.05$ level of statistical significance and the results were as follows:

Table 4. Stationarity Test Results.

VARIABLE	FORM	ADF TEST STATISTIC	P-VALUE	RESULT
<i>ITburden</i>	Original	-3.8834	0.02281	Stationary
	$f(X) \sim I(1)$	-4.5174	> 0.01	Stationary
<i>GovSize</i>	$f(X) \sim I(1)$	-3.3411	0.07728	Not stationary
	$f(X) \sim I(1)$	-4.1958	0.01038	Stationary
<i>ProxyUnem</i>	Original	-3.1105	0.1324	Not stationary
	$f(X) \sim I(1)$	-4.988	> 0.01	Stationary
<i>TimeDepIR</i>	Original	-2.7249	0.2857	Not stationary
	$f(X) \sim I(1)$	-4.6105	> 0.01	Stationary
<i>nIS</i>	Original	-2.6418	0.3188	Not stationary
	$f(\eta) \sim I(1)$	-5.1547	> 0.01	Stationary

<i>Urban_nIS</i>	Original	-2.4399	0.402	Not stationary
	$f(\eta) \sim I(1)$	-3.2699	0.09255	Not stationary
<i>Rural_nIS</i>	Original	-3.1199	0.1374	Not stationary
	$f(\eta) \sim I(1)$	-3.2615	0.09377	Not stationary
<i>Manufacturing_nIS</i>	Original	-3.2685	0.09312	Not stationary
	$f(\eta) \sim I(1)$	-11.444	> 0.01	Stationary
<i>Construction_nIS</i>	Original	-4.4116	> 0.01	Stationary
	$f(\eta) \sim I(1)$	-9.4751	> 0.01	Stationary
<i>TradeHospitality_nIS</i>	Original	-4.2532	0.01198	Stationary
	$f(\eta) \sim I(1)$	-4.7524	> 0.01	Stationary
<i>TransComms_nIS</i>	Original	-2.8944	0.2265	Not stationary
	$f(\eta) \sim I(1)$	-4.2457	0.01257	Stationary
<i>CSP_nIS</i>	Original	-3.0312	0.1734	Not stationary
	$f(\eta) \sim I(1)$	-4.9223	> 0.01	Stationary
<i>GDPpc</i>	Original	-3.3292	0.07909	Not stationary
	$f(Y) \sim I(1)$	-4.4239	> 0.01	Stationary
<i>M1</i>	Original	-4.0878	0.01443	Stationary
	$f(Y) \sim I(1)$	-4.9591	> 0.01	Stationary
<i>currentGDP</i>	Original	-3.2638	0.08905	Not stationary
	$f(Y) \sim I(1)$	-3.872	0.02354	Stationary

Based on Table 4, some of the variables under study were found to be integrated of order $I(0)$, some integrated of order $I(1)$, and some integrated of an order higher than $I(1)$. Normality was checked for at $\alpha = 0.05$ level of statistical significance, and the results were as follows:

Table 5. Normality Test Results for Individual Variables.

VARIABLE	W TEST STATISTIC	P-VALUE	RESULT
<i>ITburden</i>	0.83123	1.034e-05	No normality
<i>GovSize</i>	0.98855	0.9267	Has normality
<i>ProxyUnem</i>	0.68288	1.098e-08	No normality
<i>TimeDepIR</i>	0.88512	0.0002897	No normality
<i>nIS</i>	0.68664	1.27e-08	No normality
<i>Urban_nIS</i>	0.84158	0.0001226	No normality
<i>Rural_nIS</i>	0.81158	2.788e-05	No normality
<i>Manufacturing_nIS</i>	0.61336	2.924e-08	No normality
<i>Construction_nIS</i>	0.26302	6.549e-12	No normality
<i>TradeHospitality_nIS</i>	0.70157	5.204e-07	No normality
<i>TransComms_nIS</i>	0.1973	1.885e-12	No normality
<i>CSP_nIS</i>	0.69069	3.554e-07	No normality
<i>M1</i>	0.90602	0.001285	No normality
<i>GDPpc</i>	0.97599	0.4527	Has normality
<i>currentGDP</i>	0.95631	0.08233	Has normality

The test results from Table 5 indicate that majority of the variables in their original form do not follow the normal distribution. Most of the variables were found to not follow a normal distribution, so there was no need to test for multivariate normality.

Based on the results in Table 4, variables were transformed as follows:

$$X_1 = f(ITburden) \sim I(1)$$

$$X_2 = f(GovSize) \sim I(1)$$

$$X_3 = f(ProxyUnem) \sim I(1)$$

$$X_4 = f(TimeDepIR) \sim I(1)$$

$\eta = f(nIS) \sim I(1)$ or $f(\underline{\eta}) \sim I(1)$ where $\underline{\eta}$ is any location or industry group

$Y_1 = f(GDPpc) \sim I(1)$

$Y_2 = f(M1) \sim I(1)$

$Y_3 = f(currentGDP) \sim I(1)$

The variance inflation factors for the group of cause variables were obtained using the *car* (Fox and Weisberg, 2019) package. The results were as follows:

$$VIF_{X_1} = 1.004765 \quad \therefore \quad VIF_{X_1} < 3.3$$

$$VIF_{X_2} = 1.019224 \quad \therefore \quad VIF_{X_1} < 3.3$$

$$VIF_{X_3} = 1.156838 \quad \therefore \quad VIF_{X_1} < 3.3$$

$$VIF_{X_4} = 1.160622 \quad \therefore \quad VIF_{X_1} < 3.3$$

From the results, no multicollinearity was concluded.

Data Validation Results Summary

The data, whether split into the 1984-2019 or 1986-2019 periods or kept as the 1975-2019 period was found to generally be integrated at order $I(1)$. No multicollinearity was identified among the cause variables. The data was found to generally not follow a normal distribution; therefore, least-squares estimation would be suitable as the method of parameter estimation.

4.3 Parameter Estimation

Parameter estimation was conducted using the Ordinary Least Squares estimation method, having sub-equations be calculated in sequence. Upon testing models against the MIMIC model assumptions which were

$$\begin{aligned} E(\xi) = 0 &\sim N(0, \Psi) & \text{and} & \quad E(\mathbf{U}) = 0 \sim N(0, \sigma^2 \mathbf{I}) \\ \xi &\sim I(0) & \text{and} & \quad \mathbf{U} \sim I(0) \\ E(\xi | \mathbf{U}) &= 0 \end{aligned}$$

In the two forms of the full model, each estimated parameter was tested for statistical significance at $\alpha = 0.05$ level of statistical significance and the results were as shown in

Table 6 and Table 7. In this model, change in the growth rate of indirect tax was the only statistically significant parameter. It is also noteworthy that apart from $\hat{\lambda}_{f_1}$, the standard errors of the second form were generally lower.

Table 6. Statistical Significance and Precision of Estimated Parameters in the Full Model Based on the Informal Trade & Hospitality Industry.

ESTIMATE	STANDARD ERROR	t_{cal}	P-VALUE	RESULT
$\hat{\beta}_0 = 0.0065$	0.0417	0.155	0.878	Not statistically significant
$\hat{\beta}_1 = \mathbf{1.4738}$	0.2036	7.241	7e-08	Statistically significant
$\hat{\beta}_2 = -0.4307$	0.6386	-0.675	0.505	Not statistically significant
$\hat{\beta}_3 = 0.0373$	0.2451	0.152	0.880	Not statistically significant
$\hat{\beta}_4 = -0.0413$	0.0966	-0.428	0.672	Not statistically significant
$\hat{\lambda}_0 = -0.0083$	0.0250	-0.333	0.741	Not statistically significant
$\hat{\lambda}_{f_1} = 0.0923$	0.0669	1.380	0.177	Not statistically significant

Table 7. Statistical Significance and Precision of Estimated Parameters in the Full Model Based on the Informal Community, Social & Personal Services Industry.

ESTIMATE	STANDARD ERROR	t_{cal}	P-VALUE	RESULT
$\hat{\beta}_0 = 0.0011$	0.0256	0.042	0.966	Not statistically significant
$\hat{\beta}_1 = \mathbf{0.9359}$	0.1249	7.495	4e-08	Statistically significant
$\hat{\beta}_2 = -0.2629$	0.3917	-0.671	0.508	Not statistically significant
$\hat{\beta}_3 = 0.0756$	0.1504	0.503	0.619	Not statistically significant
$\hat{\beta}_4 = -0.0194$	0.0592	-0.328	0.746	Not statistically significant
$\hat{\lambda}_0 = -0.0080$	0.0254	-0.316	0.754	Not statistically significant
$\hat{\lambda}_{f_1} = 0.1025$	0.1083	0.947	0.351	Not statistically significant

In the two forms of the reduced model, statistical significance of the model parameters was tested at $\alpha = 0.05$ level of significance; the results are contained in Table 8 and Table 9.

Table 8. Statistical Significance and Precision of Estimated Parameters in the Reduced Model Based on the Informal Trade & Hospitality Industry.

ESTIMATE	STANDARD ERROR	t_{cal}	P-VALUE	RESULT
$\hat{\beta}_0 = 0.0063$	0.0410	0.153	0.880	Not statistically significant
$\hat{\beta}_1 = 1.4724$	0.1999	7.366	4e-08	Statistically significant
$\hat{\beta}_2 = -0.425$	0.6266	-0.678	0.503	Not statistically significant
$\hat{\beta}_4 = -0.0466$	0.0887	-0.525	0.604	Not statistically significant
$\hat{\lambda}_0 = -0.0083$	0.0250	-0.333	0.741	Not statistically significant
$\hat{\lambda}_{f_1} = 0.0923$	0.0669	1.380	0.177	Not statistically significant

Table 9. Statistical Significance and Precision of Estimated Parameters in the Reduced Model Based on the Informal Community, Social & Personal Services Industry.

ESTIMATE	STANDARD ERROR	t_{cal}	P-VALUE	RESULT
$\hat{\beta}_0 = 0.0007$	0.0252	0.026	0.979	Not statistically significant
$\hat{\beta}_1 = 0.1231$	0.1249	7.578	2e-08	Statistically significant
$\hat{\beta}_2 = -0.2512$	0.3860	-0.651	0.520	Not statistically significant
$\hat{\beta}_4 = -0.0300$	0.0546	-0.549	0.587	Not statistically significant
$\hat{\lambda}_0 = -0.0080$	0.0254	-0.316	0.754	Not statistically significant
$\hat{\lambda}_{f_1} = 0.1025$	0.1083	0.947	0.351	Not statistically significant

$\hat{\lambda}_0$ and $\hat{\lambda}_{f_1}$ estimates for both forms of the reduced model are the same as the ones in the full model in both cases. When comparing the full model and reduced model, the first form of the reduced model had lower standard errors of the estimated parameters than the first form of the full model.

A key observation from the process of model fitting before arriving at the models that satisfy model assumptions is that the causal parameters only change when any of the cause variables are added to or omitted from the model; they are not affected when indicator variables are added or omitted. However, the indicator parameters stay the same

irrespective of whether causal variables are added or not because they depend on the factor and not on the causal variables directly.

The relationship between the informal sector and GDP per capita was a growth relationship; meaning, though the effect was not statistically significant in all the models, growth in the informal sector, particularly in the Trade & Hospitality and Community, Social and Personal Services industries, had a positive effect on growth in GDP per capita. It actually could contribute to the value that the average Kenyan citizen can bring to the economy.

4.4 Goodness of Fit

4.4.1 Cointegration Test Results

Due to the fact that one of the model assumptions was $\xi \sim I(0)$ and $U \sim I(0)$, the full and reduced models based on Trade and Hospitality were found to be cointegrated at $\alpha = 0.05$ level of statistical significance while the full and reduced models based on Community, Social and Personal Services were found to be cointegrated at $\alpha = 0.1$ level of statistical significance. The complete results are as shown in Table 10.

Table 10. Cointegration Test Results on Model Residuals.

MODEL	ERROR	ADF TEST STATISTIC	P-VALUE
<u>Full Model Based on Trade & Hospitality</u> $(Y_1, Y_2) = f(\eta)$ $\eta = f(X_1, X_2, X_3, X_4)$	$\hat{\epsilon}$ \hat{U}	-3.5902 -5.102	0.04865 < 0.01
<u>Reduced Model Based on Trade & Hospitality</u> $(Y_1, Y_2) = f(\eta)$ $\eta = f(X_1, X_2, X_4)$	$\hat{\epsilon}$ \hat{U}	-3.5902 -5.099	0.04865 < 0.01
<u>Full Model Based on CSP Services</u> $(Y_1, Y_2) = f(\eta)$ $\eta = f(X_1, X_2, X_3, X_4)$	$\hat{\epsilon}$ \hat{U}	-3.4156 -4.6109	0.07242 < 0.01

<u>Reduced Model Based on CSP Services</u>			
$(Y_1, Y_2) = f(\eta)$	$\hat{\epsilon}$	-3.4156	0.07242
$\eta = f(X_1, X_2, X_4)$	\hat{U}	-4.6593	< 0.01

4.4.2 Statistical Significance of the Model

The level of statistical significance was set as $\alpha = 0.05$; therefore the model was statistically significant if F_{cal} , the value calculated by the author, was greater than F_{tab} , the value obtained from statistical tables. Results were as shown in Table 11.

Table 11. Model Statistical Significance Test Results.

MODEL	F_{cal}	F_{tab}	RESULT
Full model based on Trade & Hospitality	25.542	3.3594	Statistically significant
Reduced model based on Trade & Hospitality	25.9715	3.355	Statistically significant
Full model based on CSP Services	18.3338	3.3594	Statistically significant
Reduced model based on CSP Services	18.4998	3.355	Statistically significant

4.4.3 Goodness of fit statistics

The models also had the following goodness of fit statistics:

Table 12. Goodness of Fit Statistics.

MODEL	R^2	AIC	RMSE = $\sqrt{\text{Mean Squared Error}}$
Full model based on Trade & Hospitality	0.4222	0.04	0.2
Reduced model based on Trade & Hospitality	0.4225	0.04	0.1982
Full model based on CSP Services	0.5045	0.02	0.15
Reduced model based on CSP Services	0.5067	0.02	0.1490

In Table 12, the second form of the reduced model explained most of the variance in the model. Additionally, the second form of the full and reduced models had the lowest AIC. In terms of the root mean square error (RMSE), the second form of the reduced model had the lowest RMSE. However, the models did not generally explain a lot of the total variance; with the second form of the reduced model having a R^2 value of 51%.

4.4.4 Model Selection

For purposes of achieving the research objectives, the fourth model in Table 12 was selected. Firstly, it had statistical significance. Secondly, it had the lowest RMSE and the lowest AIC. Thirdly, it explained most of the total variation. It was visualized, having estimated parameters shown next to parameter standard errors in brackets, as follows:

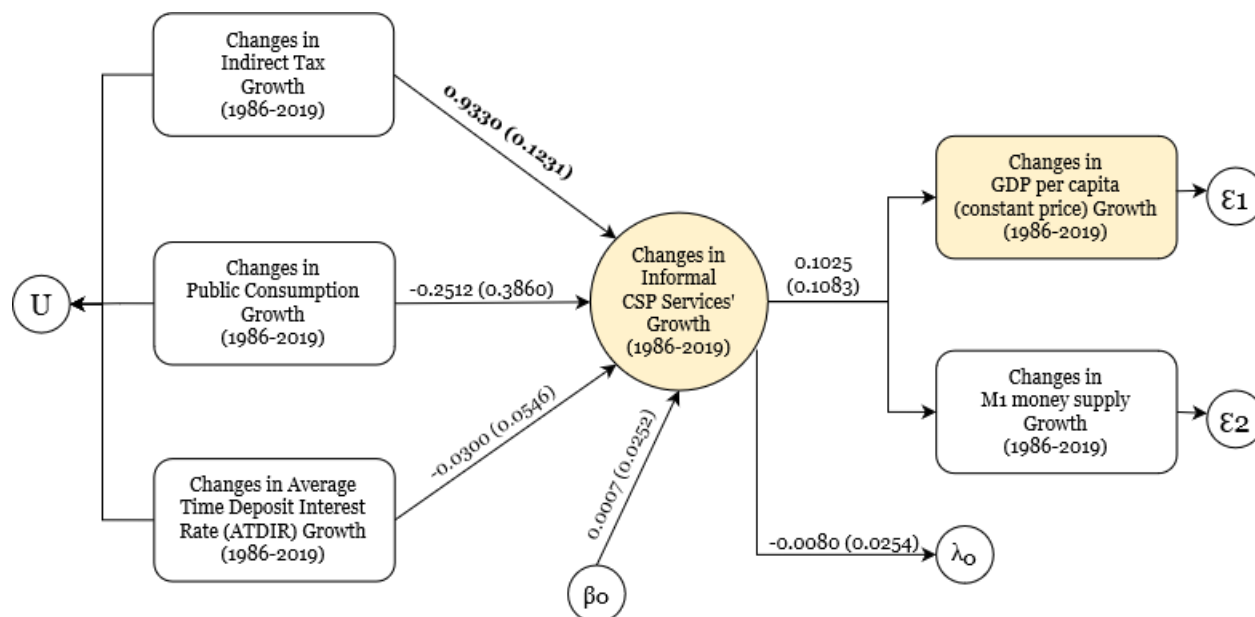


Figure 22. Reduced Model Based on the Informal Community, Social & Personal Services Industry

4.5 Model Interpretation

The various time series were found to generally be stationary, without autocorrelation. That said, the only combination of the selected variables that satisfied stationarity and normality of errors was the first two indicator variables, any of two industries in the informal sector and all cause variables. The model selected was the reduced model based on Community, Social and Personal Services informal industry, a 3-1-2 model.

This model was statistically significant, with 51% of total variation explained by the model. However, changes in growth of indirect taxes had the only statistically significant parameter. That said, changes in the growth of CSP Services had a positive effect on changes in the growth of GDP per capita at constant prices.

A unit change in the annual growth rate of the informal sector increases the growth rate of GDP per capita at constant prices and the M1 money supply growth rate by 10%.

A unit change in indirect tax annual growth rates increases the informal sector growth rate by 93%; this is a very high effect. On the other hand, changes in public consumption growth rate and time deposit growth rate only reduced informal sector growth by 25% and 3% respectively.

5 Conclusion

5.1 Summary of Results

The main objective of this paper has been to determine the type of relationship that GDP per capita and the informal sector have with each other over a time period of 34 years within the Kenyan economy.

It was found that growth in indirect taxes controlling for other economic indicators has a positive effect on growth of the informal sector, which in turn had a positive effect on growth in GDP per capita. However, the effect of the latent variable on GDP per capita growth was not a statistically significant one.

From both industry and location perspectives, the informal sector was found to have statistically significant differences between the groups at a 95% confidence level. Therefore, the informal sector can be grouped based on industry or location, and these groups' growth rates can be modelled together (forming a matrix representing the informal sector) or separately (forming a vector representing the informal sector).

Increasing indirect tax, public consumption, unemployment and time deposit growth rates at the same time by 1 produces a net effect of +103.91% on Trade & Hospitality and a net effect of +72.92% on Community, Social & Personal Services industries within the informal sector.

However when the model is reduced, increasing indirect tax, public consumption and time deposit growth rates at the same time by 1 produces a net effect of +100.08% on Trade & Hospitality and a net effect of +65.18% on Community, Social & Personal Services industries within the informal sector. The reduced model for the CSP Services industry was found to have the best fit out of all four and was selected as the preferred model; it explained 51% of total variation.

Confounding cause variables not accounted for by the model had a positive effect on the growth of the informal sector over time and the informal sector growth rate had a negative effect on confounding indicator variables. In both cases, the effects were very negligible.

The model was statistically significant in all four forms at a 95% confidence level; therefore, this relationship is a valid one.

5.2 Comparison & Contrast of Results

The results tie in with Giles' (1999) findings that taxes had a statistically significant effect. Furthermore, the variables were converted into stationary forms just like Giles (1999) did. However, in the case of this report, more years were studied, growth rates were used instead of ratios to GDP and GDP per capita was used for modelling purposes. The RMSR in Giles' (1999) best model was lower than the RMSE that the author obtained in the best model; RMSR and RMSE mean the same.

In the Medina and Schneider (2018) study, there was a mixture of use of growth rate and original variables; however, this study used strictly growth rates, with the GDP per capita held to constant prices. Additionally, the relationship may not be directly comparable, as their relationship was $\frac{\text{Informal Sector}}{\text{GDP}}$ negatively affecting growth in GDP per capita; the relationship of this paper is growth in the number of people working in informal firms positively affecting growth in GDP per capita at constant prices.

Medina and Schneider (2018) advise that their causal variables cannot be reused in other studies as the coefficients are only relative weights making a statistically significant contribution to the overall variance; consequently, the model in this report was constructed using variables derived from government sources.

Another difference between this study and the Medina and Schneider (2018) study was that they used maximum likelihood to estimate the parameters, but this study used ordinary least squares in sequence as the parameter estimation method. Finally, their estimates for the informal sector were derived but the estimates for the informal sector used in this study were obtained from annual government surveys.

Table 13 gives a comparison between the authors who used MIMIC models and the author of this report.

Table 13. Comparison of Various Author Findings.

AUTHOR	CAUSE VARIABLES	$\frac{\sum \gamma^*}{\sum \gamma}$	INDICATOR VARIABLES
Barbosa et al. (2013)	$\frac{\textit{Government Employment}}{\textit{Labour Force}}$ $\frac{\textit{Tax Burden}}{\textit{GDP}}$ $\frac{\textit{Subsidies}^*}{\textit{GDP}}$ $\frac{\textit{Social Benefits}}{\textit{GDP}}$ $\frac{\textit{Self-employment}}{\textit{Labour Force}}$ $\textit{Unemployment Rate}^*$	0/2 0/2 2/2 0/2 0/2 2/2	$\textit{Index of Real GDP}$ (1995 = 100) $\textit{Labour Force Participation Rate}$
Giles (1999)	\textit{AATR} \textit{AMTR} \textit{CPI}^* \textit{GST}^* $\textit{GST2}$ \textit{PUBEMP}^* \textit{REGS} \textit{RPDI} \textit{TAXC} \textit{TAXG} \textit{TAXLEG}^* \textit{TAXO}^* \textit{UN}	0/5 0/5 1/5 1/5 0/5 1/5 0/5 0/5 0/5 1/5 1/5 0/5	\textit{logGDP} \textit{MPRT} $\textit{CM3}$

Macias and Cazzavillan (2010)	<i>Inflation*</i> <i>Salaries*</i> <i>Tax Burden</i> <i>Unemployment*</i> <i>Gov Consumption</i>	7/7 7/7 0/3 7/7 4/4	<i>Real GDP</i> <i>Currency</i>
Medina and Schneider (2018)	<i>Trade Openness*</i> <i>GDP per capita*</i> <i>Unemployment Rate*</i> <i>Size of Government*</i> <i>Fiscal Freedom</i> <i>Rule of Law*</i> <i>Control of Corruption*</i> <i>Government Stability</i>	6/6 6/6 6/6 3/3 3/3 2/2 2/2 1/2	<i>Currency</i> <i>Labour Force Participation Rate</i> <i>Growth of GDP per capita</i>
The author	$f(ITburden)^* \sim I(1)$ $f(GovSize) \sim I(1)$ $f(ProxyUnem) \sim I(1)$ $f(TimeDepIR) \sim I(1)$	4/4 0/4 0/4 0/4	$f(GDPpc) \sim I(1)$ $f(M1) \sim I(1)$
statistical significance $\alpha = p\text{-value of } 0.05$ * $p\text{-value} < \alpha$			

5.3 Conclusions

5.3.1 Key Conclusions

The informal sector in Kenya is heterogeneous based on Chow's (1960) poolability test results. Therefore, η can be a matrix or a vector. Only two industries modelled separately were suitable to form the informal sector growth rate within the MIMIC model. They were Trade & Hospitality as well as Community and Social & Personal Services.

The informal labour force in Kenya is large and its growth positively affects GDP per capita growth; however, this positive effect is quite small (at 10% modelling with the CSP Services industry and at 9% modelling with the Trade & Hospitality industry) and not statistically significant. A possible reason for this is due to the fact that a high proportion of these workers are wage employees (Alter Chen, 2005), resulting in few businesses that contribute to GDP and GDP per capita by extension.

Government interventions greatly affect the informal sector, and the informal sector in turn affects the government's monetary policy. Public consumption and indirect tax growth rates both increased by 1, were found to have a net effect of +68% on the informal sector growth rate. This would indicate that indirect taxes have a very big impact on the informal sector, compared to all other elements in the model. This also shows that government actions affect and are affected by the informal sector.

5.3.2 Study Limitations

A key study limitation was lack of access to disaggregated data; although this data was for one country, it was aggregated annually, therefore there was no granularity within the data. Granularity would help capture any seasonalities within the time series and increase sample size, increasing the chances of the data following a normal distribution. Consequently, a negative mean was found for most of the variables.

Secondly, this data was collected from the Kenyan government. This means that researcher bias could influence the findings. If the data was collected directly from the field, different conclusions instead of those in the report could have been reached.

Thirdly, a limitation of this study is a focus only on the Kenyan economy. The world today has become more interconnected and the global economy affects the local economy to an extent; this has not been captured by the study as it takes the Kenyan economy to be a standalone entity.

5.4 Recommendations

5.4.1 Recommendations for Government

The government should conduct further research on specific informal sector groups as they are very distinct from each other, be they grouped by industry or location. As shown in the findings, the cause variables have a quarter less of the net effect on CSP Services industry that they have on Trade & Hospitality industry within the informal sector.

However, based on MIMIC model assumptions, location-based groupings may not be able to be used as the informal sector estimate and therefore a different kind of model may be more suitable to tie location grouping to economic indicators.

The government should review its indirect tax policy in light of the findings of this study, as growth in indirect tax positively affects growth in the number of people employed by the informal sector.

Finally, the government should find ways to raise the contribution that the informal sector makes to GDP per capita.

This could be achieved in the following ways. First, creation of an enabling environment that would encourage more informal wage workers, particularly youth and women, to climb the ladder and become informal employers. Second, helping each firm within the informal sector raise the overall quality and volume of the goods and services that they produce.

Third, influencing improvements in labour conditions within the informal sector. That way, they will increase the level of disposable income among informal sector employees, increasing savings and investments in the process.

Finally, in the era of the Internet, the government needs to provide an avenue that labour rights are protected within the segment of the informal sector participating in the online gig economy and look towards brokering mutually beneficial arrangements between the informal sector (firms and workers) and the global organizations that benefit from their labour or income.

5.4.2 Recommendations for Policy Makers

Direct foreign investment agencies and non-governmental institutions should seek to engage directly with the various groupings of informal sector workers and design tailored policy for the various groups that operate within the informal sector. This would be the best way to mitigate any unfair practices within the informal sector and pave the

way for meaningful work that is well compensated to increase within the informal sector and facilitate informal sector employers to increase the value of their capital assets and products.

According to the research, time deposit interest rate growth reduces informal sector growth. Therefore, impact metrics around banking the unbanked should not only consider size of loans, but also size of deposits in order to have a holistic approach to financial literacy within the informal sector.

Indirect taxes greatly impact the informal sector; therefore engagement with government to give informal firms incentives for value addition and encourage local manufacture of components could minimize the amount of indirect tax that these firms would incur due to importation of items that could easily be locally produced.

Policy makers should collectively work towards moving the conversation around the informal sector from a place of poverty and illegality, to a place of productivity and economic value.

5.5 Future Research

From an informal labour perspective, working conditions within the informal sector is an area that could be studied, as it impacts the quality of life that informal sector workers have.

From an informal firm perspective, linkages between the informal sector and the formal sector could be studied in more depth; another area of future research would be the value of capital assets owned by the informal sector as well as the types of capital assets that would be most beneficial to informal firms.

References

- Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, 591–605.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611.
- Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, 979–1001.
- ILO. (1972). *Employment, incomes and equality: a strategy for increasing productive employment in Kenya*. International Labour Office.
- Kenya National Bureau of Statistics, G. o. K. (1973). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1974). *Economic survey*. Government Printer.
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a), 631–639.
- Kenya National Bureau of Statistics, G. o. K. (1975). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1976). *Economic survey*. Government Printer.
- Cochran, W. G. (1977). *Sampling techniques* (Third edition). John Wiley & Sons.
- Kenya National Bureau of Statistics, G. o. K. (1977). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1978). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1979). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1980). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1981). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1982). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1983). *Economic survey*. Government Printer.
- Frey, B. S., & Weck-Hanneman, H. (1984). The hidden economy as an 'unobserved' variable. *European economic review*, 26(1-2), 33–53.
- Kenya National Bureau of Statistics, G. o. K. (1984). *Economic survey*. Government Printer.
- Hart, K. (1985). The informal economy. *Cambridge Anthropology*, 54–58.
- Kenya National Bureau of Statistics, G. o. K. (1985). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1986). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1987). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1988). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1989). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1990). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1991). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1992). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1993). *Economic survey*. Government Printer.

- Kenya National Bureau of Statistics, G. o. K. (1994). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1995). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1996a). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1996b). *Kenya Population Census 1989, Analytical Report: Labour Force* (Vol. 9). Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1997). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (1998). *Economic survey*. Government Printer.
- Giles, D. E. (1999). Measuring the hidden economy: Implications for econometric modelling. *Economic Journal*, F370–F380.
- Kenya National Bureau of Statistics, G. o. K. (1999). *Economic survey*. Government Printer.
- Charmes, J. (2000). The contribution of informal sector to GDP in developing countries: assessment, estimates, methods, orientations for the future.
- Kenya National Bureau of Statistics, G. o. K. (2000). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (2001). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (2002a). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (2002b). *Kenya Population Census 1999, Analytical Report: Labour Force* (Vol. 9). Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (2003). *Economic survey*. Government Printer.
- Rencher, A. C. (2003). *Methods of multivariate analysis* (Vol. 492). John Wiley & Sons.
- Bigsten, A., Kimuyu, P., & Lundvall, K. (2004). What to do with the informal sector? *Development Policy Review*, 22(6), 701–715.
- Kenya National Bureau of Statistics, G. o. K. (2004). *Economic survey*. Government Printer.
- Mitullah, W. V. (2004). A review of street trade in Africa.
- Alter Chen, M. (2005). *Rethinking the informal economy: Linkages with the formal economy and the formal regulatory environment*. WIDER Research Paper.
- Kenya National Bureau of Statistics, G. o. K. (2005). *Economic survey*. Government Printer.
- Schneider, F. (2005). Shadow economies around the world: what do we really know? *European Journal of Political Economy*, 21(3), 598–642.
- Kenya National Bureau of Statistics, G. o. K. (2006). *Economic survey*. Government Printer.
- Georgiou, G. (2007). *Measuring the Size of the Informal Economy: A Critical Review* (Working Paper No. 1). Central Bank of Cyprus.
- Kenya National Bureau of Statistics, G. o. K. (2007). *Economic survey*. Government Printer.
- Ouma, S., Njeru, J., Kamau, A., Khainga, D., & Kiriga, B. (2007). Estimating the size of the underground economy in Kenya. *KIPPRA Discussions Paper Series DP/82*.
- Kenya National Bureau of Statistics, G. o. K. (2008). *Economic survey*. Government Printer.
- Krishnakumar, J., & Nagar, A. L. (2008). On exact statistical properties of multidimensional indices based on principal components, factor analysis, MIMIC and structural equation models. *Social Indicators Research*, 86(3), 481–496.
- Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.
- Gujarati, D., & Porter, D. (2009). *Basic Econometrics* (5th edition). McGraw-Hill Education.
- Kenya National Bureau of Statistics, G. o. K. (2009). *Economic survey*. Government Printer.

- Gulzar, A., Junaid, N., & Haider, A. (2010). What is hidden, in the hidden economy of Pakistan? Size, causes, issues, and implications. *The Pakistan Development Review*, 665–704.
- Kenya National Bureau of Statistics, G. o. K. (2010). *Economic survey*. Government Printer.
- Macias, J. B., & Cazzavillan, G. (2010). Modeling the informal economy in Mexico. a structural equation approach. *The Journal of Developing Areas*, 345–365.
- Kenya National Bureau of Statistics, G. o. K. (2011). *Economic survey*. Government Printer.
- Klarić, V. (2011). Estimating the size of non-observed economy in Croatia using the MIMIC approach. *Financial theory and practice*, 35(1), 59–90.
- Yusuff, O. S. (2011). A theoretical analysis of the concept of informal economy and informality in developing countries. *European Journal of Social Sciences*, 20(4), 624–636.
- Charmes, J. (2012). The informal economy worldwide: Trends and characteristics. *Margin: The Journal of Applied Economic Research*, 6(2), 103–132.
- Kenya National Bureau of Statistics, G. o. K. (2012). *Economic survey*. Government Printer.
- Barbosa, E., Pereira, S., Brandão, E., et al. (2013). The shadow economy in Portugal: an analysis using the MIMIC model. *School of Economics and Management Working Papers*, 1–46.
- Kenya National Bureau of Statistics, G. o. K. (2013). *Economic survey*. Government Printer.
- Odera, L. C. (2013). The role of trust as an informal institution in the informal sector in Africa. *Africa Development*, 38(3-4), 121–146.
- Ogbuabor, J. E., & Malaolu, V. (2013). Size and causes of the informal sector of the Nigerian economy: Evidence from error correction mimic model. 4(1), 85–103.
- Rotich, I. (2013). *An assessment of street hawkers response to new market sites in Eldoret town, Kenya* (Doctoral dissertation). Moi University.
- Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th edition). Cengage learning.
- Bargain, O., & Kwenda, P. (2014). The informal sector wage gap: New evidence using quantile estimations on panel data. *Economic Development and Cultural Change*, 63(1), 117–153.
- Kenya National Bureau of Statistics, G. o. K. (2014). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (2015). *Economic survey*. Government Printer.
- Nchor, D., & Adamec, V. (2015). Unofficial economy estimation by the MIMIC model: The case of Kenya, Namibia, Ghana and Nigeria. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 63(6), 2043–2049.
- Posey, C., Roberts, T. L., Lowry, P. B., & Bennett, R. J. (2015). Multiple indicators and multiple causes (mimic) models as a mixed-modeling technique: A tutorial and an annotated example. *Communications of the Association for Information Systems*, 36(1), 11.
- Di Zio, M., Fursova, N., Gelsema, T., Gießing, S., Guarnera, U., Petrauskienė, J., Quensel-von Kalben, L., Scanu, M., ten Bosch, K., van der Loo, M., et al. (2016). Methodology for data validation 1.0. *Essnet Validat Foundation, Brussels, Belgium*, 1–76.

- Edmonds, W. A., & Kennedy, T. D. (2016). *An applied guide to research designs: Quantitative, qualitative, and mixed methods*. Sage Publications.
- Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd edition). Sage Publications.
- Kenya National Bureau of Statistics, G. o. K. (2016). *Economic survey*. Government Printer.
- Nordman, C. J., Rakotomanana, F., & Roubaud, F. (2016). Informal versus formal: A panel data analysis of earnings gaps in Madagascar. *World Development*, 86, 1–17.
- Kenya National Bureau of Statistics, G. o. K. (2017a). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (2017b). *Kenya Population and Housing Census 2009, Single and Grouped Ages in Years by County and District*. <https://www.knbs.or.ke/?wpdmpro=single-and-grouped-ages-in-years-by-county-and-district>
- Kenya National Bureau of Statistics, G. o. K. (2018). *Economic survey*. Government Printer.
- Meagher, K. (2018). Cannibalizing the informal economy: Frugal innovation and economic inclusion in Africa. *The European Journal of Development Research*, 30(1), 17–33.
- Medina, L., & Schneider, F. (2018). Shadow economies around the world: what did we learn over the last 20 years?
- Njangang, H., Noubissi, E., Nkengfack, H., et al. (2018). Do remittances increase the size of the informal economy in Sub-saharan African countries?" *Economics Bulletin*, 38(4), 1997–2007.
- Benanav, A. (2019). The origins of informality: the ILO at the limit of the concept of unemployment. *Journal of Global History*, 14(1), 107–125.
- Dragsted, B. (2019). Crackdown economics: Policing of hawkers in Nairobi as violent inclusion. *Geoforum*, 102, 69–75.
- Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third edition). Sage. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Kenya National Bureau of Statistics, G. o. K. (2019). *Economic survey*. Government Printer.
- Kenya National Bureau of Statistics, G. o. K. (2020). *Economic survey*. Government Printer.
- Tonuchi, E., Idowu, P., Adetoba, O., & Mimiko, D. (2020). How large is the size of Nigeria's informal economy? A MIMIC approach. *International Journal of Economics, Commerce, and management*, 8(7), 204–227.
- Trapletti, A., & Hornik, K. (2020). *tseries: Time Series Analysis and Computational Finance* [R package version 0.10-48.]. <https://CRAN.R-project.org/package=tseries>
- R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>