Public Investment and Inequality: Panel evidence for South Africa

Mathane Makgatho*†

August 30, 2017

Abstract

This paper uses panel data of nine South African provinces from 1995-2013 to decompose the impact of public investment on provincial inequality. I consider public investments in the form of education, health, water, sanitation, agriculture and electricity. The results suggests that private capital and labour are positive and significant. Inputs that relate to public investments that are positive are the electricity utility company's capital expenditure roll-out, provision of water and sanitation and education expenditure. The estimated coefficients are used to decompose the effect of public investment on provincial inequality using the Fields and Shapley method. The two approaches suggest that private capital and labour explain most of the inequality with Fields approach suggesting that the two contribute over three quarters of provincial inequality while the Fields method suggest that the two variables contribute 43 percent of provincial inequality. The Fields methods suggest that education expenditure has a neutralising effect while the electricity utility company's capital expenditure, provision of water and sanitation as well as support of agricultural sector have an inequality reducing effect. The Shapley method on the other hand, suggest that education expenditure, electricity utility company's capital expenditure, provision of water and sanitation as well as support of agricultural sector have a neutralising effect on provincial inequality.

*PhD Student, Department of Economics, Pretoria, South Africa, email: mathane@matlacapital.co.za

†Supervisor - Dr Yoseph Getachew, -Department of Economics - University of Pretoria, Pretoria, South Africa,

1 Introduction

The role of public investment as a key driver and determinant of economic growth has long been acknowledged and debated by academics and policy makers alike. The speed and pace of economic growth is of interest as sustained growth can raise per capita income levels and assist in reducing levels of poverty and inequality. According to Barro (2004), understanding the determinants of growth is crucial to understanding the drivers increasing standards of living and poverty reduction. Public investment is an important policy tool that governments can use to improve welfare and equity, Martin (1999). The effectiveness of public investment as a policy tool lies in understanding the drivers behind increased standards of living and a more equal society. Public investments can improve equity by promoting growth through investments in social and economic infrastructure and indirectly through creating an enabling environment for the private sector to thrive. These direct and indirect benefits of public investments are expected to increase income and improve well being of a nation.

Between 1996 and 2013, the South African economy grew by an average of 3.2 percent. Despite this steady albeit low growth as reflected by the graph below, South Africa continue to be an unequal society as reflected by the Gini coefficient. In fact, inequality increased from 0.63 in 1995 to 0.69 in 2014 (Quantec dataset) despite the country recording a positive average growth rate for the same period. The first ten years of democracy saw a huge increase of the gini coefficient between 1995 and 2005 before declining though in 2013, it is still higher than in 1995. The same period, saw massive increases in social security spending coupled with increases in both health and education spending as well as emphasis on state led infrastructure investment.

For the same period, the nine provinces all recorded a growth though at differing magnitudes. The two richest provinces, the Western Cape and Gauteng both recorded 3.6 percent average growth which is above the national average. Whereas provinces like North West, Free State and Northern Cape recorded 2.0; 2.1 and 2.2 percent average growth. In the same way, as at national level, inequality increased across all nine provinces with Western Cape leading with 12.8 percent increase and Gauteng increased by 9 percent. The lowest increase was recorded by Limpopo at 1.7 percent followed by Mpumalanga at 2.9 percent.

Figure 1: Gini coefficient 1995-2013

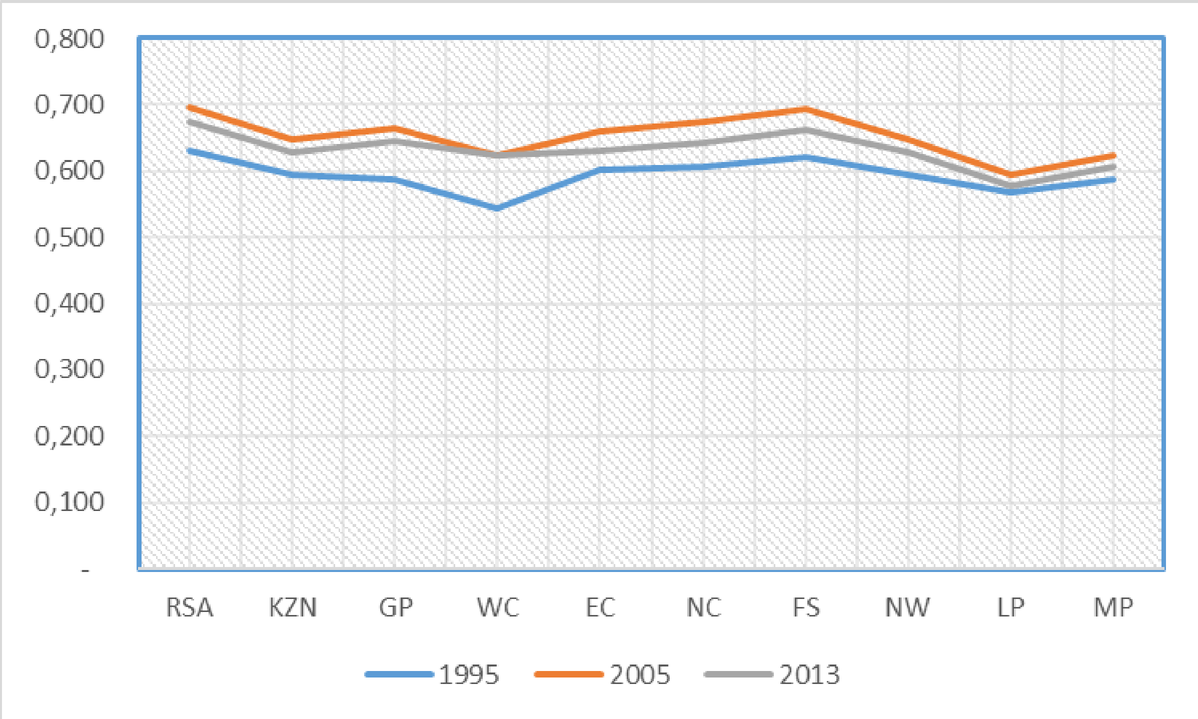

Figure 1 shows that inequality remains high and increased for all nine provinces in the last two decades despite increased public investment. Given the above data and trend, it appears that public investment as a tool has not been effective in achieving the desired results.

The purpose of this paper is to analyse how public investment impact provincial inequality by combining growth theory with economic geography when increasing returns are present. This is because determinants of provincial economic growth and inequality are much more complex than determinants of national growth and inequality. To understand provincial inequality, one must analyse the determinants/drivers of provincial growth and how economies of scale interact with transportation costs. Despite several studies on the effects of infrastructure on growth and inequality, most studies have either focused at the national level using aggregate investment or the local level using only a subset of provincial/regional data. Not many studies have tried to unpack the impact of different types of infrastructure on inequality and employment. Zhang and Fan (2004) pointed out that different types of public goods have different externalities which may cause differing impact on growth, employment and equity. There are limited studies that focused on provinces as most international studies focused on broader national issues. Further-more, not many have included the impact of geography and provincial endowments on provincial economic outcomes.

This paper contributes to the body of literature on growth, public investment and inequality by adding the effect of public investment on provincial inequality which has not been extensively studied. In this paper, i use a framework developed by Zhang and Fan (2004) to assess the impact of public investment on provincial inequality. This is based on the assumption that different types of public investment have different impact on provincial inequality. I consider public investments in the form of education, health, water, sanitation, agriculture and electricity. South African provinces are chosen for two reasons. The first one is that inequality is stubbornly high and provincial per capita income trajectory points to more divergence and pose a threat to social stability and secondly, limited government resources should be channeled to high impact public investment. We estimate a production function which includes conventional inputs and different types of public investment and use the estimated coefficient to decompose the impact of public investment on provincial inequality.

1.1 History of South Africa's Provincial inequality

Prior to democratic dispensation in 1994, South Africa was divided into four provinces and homelands which mainly separated Blacks and Whites. These provinces, Orange Free State; Transvaal; Natal and the Cape Province which were meant for Whites and urban Blacks. Blacks were given homelands per their ethnicity. For example, Qwaqwa was created for Blacks of Basotho origin and KwaZulu for a Black person of Zulu origin. The homelands were ten in total and were Qwaqwa; Lebowa; KwaZulu; KwaNdebele; KaNgwane; Gazankulu; Venda; Transkei; Bophuthatswana and Ciskei.

The four provinces with the most of the land and resources represented White South Africa while the homelands with limited resources and limited development represented Black South Africa. These were also developed with targetted industrial nodes. The homelands were not developed and Blacks relied on White South Africas economy for their survival.

Farm land was not in good condition due to overgrazing and soil erosion as

> blacks owned only thirteen percent of land. Thus, millions of Black people had to leave their homelands to work in the mines, for White farmers and other industries in the cities. The homelands served as labour reservoirs, housing the unemployed and releasing them when their labour was needed in White South Africa(sahistory.org.za).

Post 1994, the provinces were rearranged into nine new provinces without setting aside specific locations for racial groups. The former homelands ceased to exist and were incorporated into the new provinces. Out of the nine provinces, only three provinces(Gauteng; Western Cape and Northern Cape) did not have to include former homelands given their geographical boundaries. The new provinces are a mixture of diverse economic and social backgrounds given South African history and all are faced with challenges of unemployment, poverty and inequality. The government is trying to reverse inequality and poverty by using public investment as a policy tool. But given the government's fiscal constraints, limited resources need to be channeled to the most effective investment. This study aims to provide policy makers with an analysis of which public investment can be used to effectively address inequality.

2 Literature Review

In his seminal paper, Aschauer (1989) found a strong and positive relationship between infrastructure and output for USA and found that main infrastructure in the form of water, roads and power contributed more than social infrastructure. Using cross sectional country data, Easterly and Rebelo (1993) investigated different types of public investment on economic growth. Their results indicated that transport and communication have a strong relationship with growth especially when estimated using instrumental variables.

Fedderke et al. (2006) and Perkins et al. (2005) have indicated a positive and a strong relationship between infrastructure and economic growth for South Africa. The studies have in particular found that aggregate infrastructure stock and investment drive growth and the relationship depends on various types of infrastructure. These studies further found that there are a number of channels through which infrastructure affect growth. Fedderke highlights the five channels of impact as through factor of accumulation; stimulate aggregate demand and can be used as a tool for industrial policy.

On the relationship between infrastructure investment and inequality, Klenert et al. (2014) studied the trade off between equity and economic growth when public investment is financed through taxes. Using heterogeneous agents model based on differences in savings behaviour, income sources and time preferences, their results differ from Chatterjee and Turnovsky (2012) whose heterogeneity only comes from initial endowments. Their study shows that tax financed public investment does not increase inequality but has an inequality reducing effect.

In his paper, Getachew (2010) developed a theory of public capital, inequality and growth in a two sector model to explain income inequality. Assuming that agents have different initial wealth and credit market imperfection, he showed that evolution of income distribution depends on shares of private production factors. The more the initial wealth, the more likely households will be able to tap into investment opportunities which are available to those with more initial wealth as credit markets are assumed to be imperfect. On the other hand, using a model of heterogeneous endowments within an economy subject that is subject to idiosyncratic production shocks, Getachew and Turnovsky (2015) showed the relationship depends on elasticity of substitution between private and public capital.

Demurger (2001) investigated the relationship between differences in provincial infrastructure and economic growth using a panel of 24 Chinese provinces from 1985 -1998. He estimated a growth equation using Barro type model and accounted for differences in economic environment, physical capital, human capital as well as took geography and infrastructure endowment into account. The results indicate that differences in location, transport infrastructure and telecommunications do account for variation in economic performances across provinces. Demurger indirectly combines growth theory with economic geography without explicitly linking the behaviour of firms and consumers to availability of infrastructure and location choice. The model does not address how availability of infrastructure and telecommunications contributes to provincial growth differences.

On the empirical side, Zhang and Fan (2004) used provincial level data for rural China to show the impact on regional inequality of different types of public investments. They showed that regional variations on inequality are large and that different types of investments have different impact. Estimating the impact of rural electrication on employment, Dinkelman (2011) found that rural electrication has a positive effect on labour markets and that the positive impact on women is more pronounced.

? investigated the link between physical infrastructure and income inequality using both cross country and panel data econometric techniques. The paper used telecommunications, energy, roads and rail track as variables for infrastructure and initial GDP per capita, average annual GDP growth and secondary school enrollment as control variables. The results of the study suggest that there is a strong link and negative link between quantity of infrastructure and income distribution. This impact appears to be much stronger in poor countries rather than the rich countries. But, inequality of infrastructure seems to be more important in rich countries and less so in poor countries.

Models that include public infrastructure in the production function has been criticized by Martin and Rogers (1995) and Krugman (1991) as failing to address the role played by public infrastructure in regional interaction. They advocate for models that combines economic geography with increasing returns and model infrastructure as intervention to the relationship between the firms and consumers with main impact being through demand. These types of models are useful in answering questions relating to why firms choose specific locations as opposed to the alternative.

In his paper, Krugman (1991) combines economic geography which studies the location of factors of production with endogenous model. Using a simple two region model with two kinds of production: agriculture and manufacturing he models endogenous growth with economic geography. Agriculture activities are tied to land and is a constant returns sector whereas manufacturing is an increasing returns sector which can be located in either of the two regions. Assuming pecuniary externalities associated with either demand or supply linkages, he shows that with lower transportation costs and strong economies of scale, manufacturing will concentrate in whichever region that gets a head start. His analysis and conclusion elevates initial conditions of regions as key determinant of what is likely to happen in the future. He links initial conditions to firms choice of location which in turn is dependent on availability of market, labour and transportation costs.

In the spirit of Krugman, Martin and Rogers (1995) analysed the effect of public infrastructure on industrial location when increasing returns are present. They also used a two region model to assess different types of infrastructure. Unlike the Krugman model, their model differentiates between public infrastructure that improves domestic trade. They assume fixed capital requirement for each differentiated good and also assumes internationally immobile labour but inter-sectoraly mobile labour. These assumptions allow labour to move freely between sectors depending on the wage rate and income prospects, while international immobility of labour captures what happened to labour income and labour prospects if firms choose a specific location that is out of reach for local labour. The same assumptions can be valid for regional dynamics in a similar manner as international dynamics.

The model offers interesting insights as it includes capital endowment differences which captures heterogeneity of regions. They assert that a country/region that is initially rich in capital will attract more capital. This is based on the observation that regions with better international infrastructure magnify differentials in capital endowment. The paper concludes by asserting that for convergence to be achieved, policy should favour infrastructure that improves and facilitates domestic trade rather than international trade. If we use this conclusion for regional infrastructure, this implies that for regional convergence to be achieved, policy should be biased towards infrastructure that promotes trade within regions as opposed to trade between regions.

Using quantile regression method for Mexico that assesses impact of public investment across income distributions, Costa-i Font and Rodriguez-Oreggia (2005) found that returns to public investment differs according to regional income distribution. The paper also found that initial GDP per capita is positive and significant for all quantiles implying that there is some path dependency in the evolution of regional income i.e. those regions that are initially rich will get more returns from public capital. These results suggest heterogeneous regional impact from public investment and that public investment has no particular impact on low income regions and are therefore ineffective in reducing regional inequalities. But, for richer regions, public investment causes those regions to be even more richer due to agglomeration effect. This study suggests that public investment worsens regional disparities and highlight the importance of regional integration of initially well-endowed regions.

3 Empirical Analysis

3.1 Data

We use a panel data set for all nine provinces of South Africa namely Gauteng, Free State, KwaZulu Natal, Eastern Cape, Northern Cape, Western Cape, North West, Mpumalanga and Limpopo over a period of 1995 to 2013 (T=19 and N=9). Provincial data is only available from 1995 as prior to 1994, South Africa was organised into four provinces and ten homelands. The huge differences in demarcation post 1994, makes it almost impractical to compare provincial on a like for like before 1994. Therefore, due to data constraints prior to 1994, the research period starts from 1995.

The dependent variable is log of provincial output (Iny). Independent variables are the conventional variables which are log of labour (Inemp) and log private capital (Inprivcapital) and public infrastructure variables which are represented by percentage of households with tap water in yard; education expenditure as a percentage of provincial expenditure; health expenditure as a percentage of provincial expenditure; agricultural loans disbursed by the Land bank; percentage of household with flushed toilet and capital expenditure spent by Eskom the electricity utility for power mega infrastructure. Sources of data are the National Treasury for health and education data; Land bank for agriculture loans disbursed; Eskom for capex roll-out information and the rest from Quantec dataset.

As discussed in the literature section, tax financed public investment is expected to have an inequality reducing effect and public investment in general is expected to have a strong a positive effect on growth. When public investment is disaggregated into different types of literature, human capital in the form of education and health is expected to have a positive impact on growth and electrification as highlighted in Dinkelman (2011) has a positive effect on labour market with pronounced impact on women.

To control for differences in regions, we include distance to port and distance to Johannesburg as proxies for access to international markets and access to domestic markets. We also include initial GDP to measure effects of a head start and tress index to capture impact of the structure of the provincial economies. Lastly, we include a dummy variable to capture effects of being a rich region. The three provinces, Gauteng; Western Cape and Kwazulu Natal are taken as rich with the remaining six taken as poor.

As per literature review in the earlier section, it is expected that for initial wealth will play an important role in wealth creation. The more the initial wealth the more households are able to tap into investment opportunities especially when credit markets are assumed to be imperfect. Initial conditions are likely to be a key determinant of what is likely to happen in future as well location of production factors. The closer the province to domestic and international markets as proxied by distance to Johannesburg and distance to nearest port, the better of that region is expected to be.

The table in the next page shows the correlation matrix of the variables used. Private capital, employment, agriculture, the structure of the provincial economy(tressindex) and to a lesser extent education are all positively related to provincial output. Rich province indicator is all positively related to provincial output which might indicate that the sate of development and endowment matter for growth and prosperity. On the other hand, key public investment like electricity, water and health seem to have a negative correlation with provincial output.

When we look at the correlation between gini coefficient, the matrix indicates that it is negatively correlated with employment, private capital, water and health public investment. This might imply that these variables have a negative impact on inequality. Whereas, variables like education and electricity seem to have a positive impact on inequality.

	InY	EMP	PRIV	ELECTR	WATER	HEALTH	EDU	AGRIC	GINI	PERCAPITAY	TRESSINDEX
InY	1.0000										
EMP	0.8880*	1.0000									
PRIV	0.9581*	0.8402*	1.0000								
ELECTR	-0.1187	-0.2118*	-0.1212	1.0000							
WATER	-0.2310*	0.0441	-0.2916*	0.0634	1.0000						
HEALTH	-0.7413*	-0.5983*	-0.6909*	0.1856*	0.5053*	1.0000					
EDU	0.1219	-0.0831	0.1838*	0.1125	-0.3312*	-0.1616*	1.0000				
AGRIC	0.3008*	0.2034*	0.2546*	-0.0107	-0.1194	-0.3552*	0.0913	1.0000			
GINI	-0.0656	-0.1689*	-0.1186	0.1028	-0.3142*	-0.2368*	0.3116*	0.0995	1.0000		
PERCAPITAY	0.3037*	0.2506*	0.2866*	-0.1550*	-0.1001	-0.4119*	0.2150*	0.2534*	0.0715	1.0000	
TRESSINDEX	0.2150*	0.1921*	0.0053	-0.0486	0.2642*	-0.1660*	-0.0104	0.2113*	-0.0013	0.2279*	1.0000
INITGDP	-0.2807*	-0.3539*	-0.1800*	0.0000	-0.0269	0.3973*	-0.3326*	-0.1488	-0.1926*	-0.1653*	-0.3981*
D-JHB	0.0785	0.0420	0.0240	-0.0000	0.0074	-0.1720*	0.1558*	0.2263*	-0.1734*	0.0860	0.2393*
D-PORT	-0.5289*	-0.5517*	-0.4844*	0.0000	-0.0389	0.4187*	-0.2081*	-0.2040*	-0.0060	-0.0937	-0.1717*
RICHPROV	0.8172*	0.6989*	0.8563*	-0.0000	-0.3278*	-0.6825*	0.4462*	0.1744*	-0.0748	0.2813*	0.0040

3.2 Estimation strategy

Our panel data set consist of N=9 and T=19 over a period of 1995-2013 and covers all of South Africas nine provinces. We use panel data regression as this strategy allows us to study and analyse South Africa at a disaggregated level taking into account the unique features of each province without having to generalise. These types of panel data often have challenges of heterogeneity and endogeneity biases. Panel data regressions can control for heterogeneity by being able to identify specific effects that are not easily detectable using pure cross section or pure time series and are better able to study dynamics of adjustment, Baltagi (2008).

Most of the data used is non stationary and we therefore follow the approach of Phillips and Moon (1999) who investigated regressions with non stationary panel data where both T and N are large. They developed a limit theory to help understand and interpret regressions with non stationary data. They showed that for panel data observations, pooling the cross section and time observations may reduce the strong effect of the residuals while retaining the strength of the signals. Their results indicate that the challenge of spurious regression is less so in panel data because of the averaging in panel estimators and therefore pooled panel can provide consistent estimates for long run regressions.

The application of the classical linear regression requires that key assumptions are satisfied. In practice, the assumptions of the classical model do not hold. Though these assumptions have no effect on the ordinary least squares(OLS) regression technique per se, they do affect the properties of the OLS estimators as well as the resulting tests statistics. If we estimate a regression assuming that all regressors are exogenous, when in fact they are not, the estimates will not be efficient. However, this assumption will be violated if there are individual-specific or time effects and if these effects are correlated with one of the regressors. In our case, it seems there may be individual effects that have to be taken into consideration. Therefore, some alternative to the straightforward application of the classical linear model should be used.

Generalized Least Squares (GLS) is most used when there is heteroscedasticity and autocorrelation in the variables: the errors have unequal variances and/or are correlated; Var (y) is no longer a scalar variance-covariance matrix. In this case, there is no guarantee that the OLS estimator is the most efficient within the class of linear unbiased (or the class of unbiased) estimators. GLS is introduced to improve upon estimation efficiency when var(y) is not a scalar variance-covariance matrix. The GLS estimator transforms the dependent and explanatory variables and renders the OLS estimates efficient. In this case, heteroscedasticity and autocorrelation is suspected and we therefore use estimation techniques that corrects for heteroscedasticity and autocorrelation.

The issue of statistical endogeneity which is the unobserved individual effects nested in the error term might be correlated to the regressors is dealt with by using fixed effects and heterogeneity of intercepts is also dealt with using fixed effects with robust standard errors. Economic endogeneity caused by reverse causality between output and private capital is dealt with using Fixed effects with Instrumental variables. The xed effects control for heterogeneity across provinces such as geographic area, access to domestic and international markets and composition of provincial economies. Instrumental variables (IV) are used to reduce the problem of possible statistical endogeneity in the form of unobserved heterogeneity, as well as economic endogeneity in the form of reverse causality. Reverse causality may be present in the model through output and private investment. The instrumental variable method allows consistent estimation when the explanatory variables are correlated with the error terms of a regression relationship.

Lastly, we estimate a dynamic model as a number of problems may arise from a static model as Arellano and Bond (1991) highlighted in their paper. The first problem is that private capital is assumed endogenous because causality may run both directions from output to private capital and vice versa. These regressors may be correlated with the error term. The second challenge is that time invariant province specific effects may be correlated with the error term as the error term consists of unobserved province specific effects and observation specific effects as indicated byBaltagi (2008).

$$u_{it} = v_i + e_{it} \tag{1}$$

The third problem is the presence of a lagged variable that causes autocorrelation. We resolve these problems by using the Arellano Bond difference GMM to estimate the dynamic equation. We use the lagged levels of regressors as instruments which make the endogenous variables pre-determined and therefore not correlated with the error term.

We resolve the problem of time invariant charecteristics being correlated with the error term by using the first difference to transform the main equation into a general form of:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \beta \Delta x_{it1} + \Delta u_{it} \quad (2)$$

This transformation removes province specific effects that are time invariant.

$$\Delta u_{it} = \Delta v_i + \Delta e_{it} \quad (3)$$

which becomes:

$$\Delta u_{it} = \Delta e_{it} + \Delta e_{it-1} \quad (4)$$

Roodman (2006) asserts that, the Arellano Bond estimator was created for small T and large N. He suggests that in cases of a large T, a shock to the fixed effects in the error term dies out with time and the correlation between the error term and lagged dependent variable becomes minimal.

Once we have satisfied ourselves that the model is correctly specified, we use the regression results of equation 2 to decompose the level of inequality as measured by log variance of income. We use the results to compute the relative contribution of each explanatory variable to inequality for both the Shapley and Fields methods and compare the results.

3.3 The model

We follow the approach used by Zhang and Fan (2004) to unpack the determinants of provincial inequality in South Africa. The approach assumes that each region has the same production function but lie on different points on the production surface. The model is extended to include industrial location, initial conditions and structure of the provincial economy as control variables to capture the heterogeneity of regions and the interaction of firm's location decision with consumer income and demand. This is in the spirit of Krugman (1990) and attempts to address the impact of public investment in regional interaction. We start by estimating the following production function which includes conventional inputs and different types of public investment.

$$Y = A\prod_{i=1}^{k} X_i^{\beta_i} \prod_{j=1}^{m} P_j^{\gamma_j} \prod_{o=1}^{o} Q_n^{\lambda_n} \tag{5}$$

where Y is the regional output, A the intercept, X the conventional inputs, P public investments and Q as control variables. Where β, γ and λ are output elasticities with regard to conventional inputs, public investment and control variables. The error term is included to capture stochastic shock to output and is assumed to be uncorrelated to explanatory variables.

The logarithmic form of equation 1 is therefore given by:

$$y = a + \sum_{i=1}^{k} \beta_i X_i + \sum_{j=1}^{m} \gamma_j P_j + \sum_{o=1}^{o} \lambda_n Q_n + \epsilon \tag{6}$$

The first step in decomposing inequality is to estimate an income generating model. We will estimate the model in equation 2 above and use the estimated equation to decompose the impact of explanatory variables on inequality. There are a number of methods that have been widely used in regression based decomposition with the common ones being the Field method and Shapley decomposition methods. Inequality decomposition literature was spearheaded by Shorrocks (1982) and Shorrocks (1984) where he studied decomposition of inequality by income sources and population subgroups. In his 1982 work, he used the assumption that separate income source contribution sum up to the overall inequality value and demonstrated this using six assumptions. Following on Shorrocks, the variance of Y in equation 2 can be decomposed as follows:

$$\sigma^2(y) = [\sigma^2(\beta_i x_i) + \sum_{i \neq k} \rho_{i,k})\sigma(y)\sigma(x_i)] + [\sigma^2(\gamma_j p_j) + \sum_{j \neq m} \rho_{i,k})\sigma(y)\sigma(p_j)] + [\sigma^2(\lambda_n q_n) + \sum_{i \neq k} \rho_{i,k})\sigma(y)\sigma(q_i)] + [\sigma^2(\epsilon) + \sum_{rho} \sigma(\epsilon)]$$

where $\sigma^2(y)$ is the variance, with y in logarithmic form; ρ represent the correlation coefficient between y and the rest of the variables.

Per Shorrocks (1982), lets assume that:

$$\sigma^2(y^k, y) + \sum_{j \neg k} \rho_{j,k}\,\sigma(y^j \sigma(y) = cov(y^k, y) \tag{7}$$

then:

$$\sigma^2(y) = \sum_{i=1}^{k} cov(y, \beta_i x_i) + \sum_{j=1}^{m} cov(y, \gamma_j p_j) + \sum_{o=1}^{o} cov(y, \lambda_o q_o) + cov(y, \epsilon) \tag{8}$$

$$\sigma^2(y) = \sum_{i=1}^{k} \beta_i cov(y, x_i) + \sum_{j=1}^{m} \gamma_j cov(y, p_j) + \sum_{o=1}^{o} \lambda_o cov(y, q_o) + \sigma^2(\epsilon) \tag{9}$$

Contribution of factor x is given by:

$$S_x(\sigma^2) = cov(y, x) \tag{10}$$

and the proportional contribution is given by:

$$S_x(\sigma^2) = \frac{cov(y, x)}{\sigma^2(y)} \tag{11}$$

The above mentioned traditional methods have been criticized as merely descriptive and that they fail to detect and measure the contributions of individual factors and therefore of limited relevance to policy makers Manna et al. (2012) and Wan and Zhou (2005). Regression based methods are advocated where standard income-generating equations written in terms of covariances are estimated. They are able to unpack the determinants of the dependent variable as they go beyond description. These observed determinants of inequality may include economic, geography, social and policy variables.

Manna et al. (2012)further asserts that the regression based methods can manage problems of endogeneity due to reverse causality which is limited with traditional approaches. This argument is supported by Morduch and Sicular (2002)and asserts that the lack of endogeneity control is the main reason that limits the decomposition to being purely descriptive. Wan and Zhou (2005) and Heshmati (2004) also advocates for regression based approach as they can control for heterogeneity which is not possible with traditional approaches.

Modern methods includes Fields (2003)and Shorrocks (2013). This paper follow the Shapley procedure as advocated by Shorrocks. He asserts that it is an attractive procedure because it treats factors in a symmetric manner; contributions sum to amount that needs to be explained and lastly, those contributions can be interpreted as the expected marginal effects. The Shapley decomposition considers the marginal effect of eliminating explanatory variables in sequence.

4 Estimated Results

4.1 Pooled OLS

We start of by estimating a Pooled Ordinary Least Squares model (POLS) which is reported in Table 1. We start off with a basic model with the dependent variable is log of provincial output (Output) and conventional independent variables and investment on education and health. Independent variables are the conventional variables which are log of labour (Emp) and log private capital (capital) and public infrastructure variables which are represented by percentage of households with tap water in yard(water); education expenditure as a percentage of provincial expenditure(edu); health expenditure as a percentage of provincial expenditure(health); agricultural loans disbursed by the Land bank(agric); percentage of household with flushed toilet(sanitation) and capital expenditure spent by Eskom the electricity utility for power mega infrastructure(eskom). We then add other variables in a step wise fashion to check robustness of the model.

The results reports the regression results and indicate that overall, the model is a good fit as the R squared is very high at between 0,95 and 0,98. Random effects is not preferred as the data is not randomly selected but the entire population of nine provinces is considered.

Our specification includes control variables that captures differences across provinces. These are dummy variable for rich provinces, with Gauteng, Western Cape and Kwazulu Natal taken as rich and the remaining six provinces as poor. We included distance to port and distance to Johannesburg to capture potential impact of access to domestic and international markets on provincial income. Lastly, we have included a tress index as a measure of how diversified or concentrated the provincial economy is. An index closer to 100 indicates high concentration and one closer to zero indicates diversification.

All of included variables are significant and most of them have the expected signs. Both the conventional inputs of employment and the lag private capital are positive and significant and in line with theory. However, water as represented by percentage of households with tap water in yard and share of health expenditure relative to total provincial government expenditure are negative and significant. This implies that the two variables have a negative impact on provincial income. Human capital theory suggests that investments in health and education are expected to both have a positive contribution on output and income which is not the case in this instances.

Table 1: Pooled OLS

	POLS 1	POLS 2	POLS 3	POLS 4	POLS 5	POLS 6
Capital	0.702***	0.604***	0.602***	0.835***	0.740***	0.364***
Emp	0.304***	0.445***	0.453***	0.213***	0.126*	0.249***
Edu	0.038	0.070	0.046	0.265***	0.305***	0.276***
Health	-0.261***	-0.091	-0.089	-0.269***	-0.299***	-0.249***
Electricity		-0.299***	-0.319***	-0.621***	-0.756***	-0.796***
Eskom		0.375**	0.401**	0.380***	0.269*	0.320**
Agric			-0.029	-0.063***	-0.054**	-0.039*
Water				0.573***	0.458***	0.188*
Sanitation				-0.061	0.170	0.243*
Distancejhb					-0.000	-0.005*
Tressindex					-0.026	0.033*
Distanceport					-0.012**	-0.003
Initgdp1						5.713***
_cons	-24.083*	-15.983	-7.384	4.902	37.681*	-16.394
N	171	171	171	171	162	162
R-squared	0.954	0.959	0.960	0.977	0.978	0.984
Adj R-squared	0.953	0.958	0.958	0.976	0.976	0.982
F	857.396	643.764	556.052	689.108	498.848	626.123

The dependent variable is log of provincial output (Output). Independent variables are the conventional variables which are log of labour (Emp) and log private capital (capital) and public infrastructure variables which are represented by percentage of households with tap water in yard(water); education expenditure as a percentage of provincial expenditure(edu); health expenditure as a percentage of provincial expenditure(health); agricultural loans disbursed by the Land bank(agric); percentage of household with flushed toilet(sanitation) and capital expenditure spent by Eskom the electricity utility for power mega infrastructure(eskom). Sources of data are the National Treasury for health and education data; Land bank for agriculture loans disbursed; Eskom for capex roll-out information and the rest from Quantec dataset.

Distance to Johannesburg and distance to port are used as proxies for access to domestic market and access to international markets. Coefficient for both variables are negative and significant. This implies that those provinces that are further away from Johannesburg are worse off as compared to those within close proximity of Johannesburg. Close proximity to a port also appears to be an advantage over those that are further away from ports and therefore limited access to international markets. These findings tie with economic geography debate that asserts that regions that are close to markets and have better endowment tend to attract more capital and therefore be even more richer.

4.2 Fixed Effects Model

To control for heterogeneity across provinces such access to international and domestic markets, initial level of development, structure of provincial economy, we estimate the fixed effects model and fixed effects with instrumental variable to control for endogeneity. Instrumental variables (IV) are used to reduce the problem of possible statistical endogeneity in the form of unobserved heterogeneity, as well as economic endogeneity in the form of reverse causality. This methods allows for consistent estimation.These types of panel data often have challenges of heterogeneity and endogeneity biases. We include control variables that capture difference across provinces We also use fixed effects model to deal with statistical endogeneity (unobserved individual effects in the error term which might be related to repressors). To deal with problems of endogeneity, we use instrumental variable technique

Table 2: Fixed Effects Model

	FE-1	FE-2	FE-3	FE-4	FE-5	FE-6	FE-8
Capital	1.077***	0.813***	0.456***	0.412***	0.409***	0.372***	0.372***
Emp	-0.040	0.170**	0.015	0.147***	0.020	0.148**	0.148**
Edu	0.093	-0.202**	0.071	0.119*	0.037	0.109	0.109
Health	-0.314***	-0.160**	-0.025	-0.010	0.006	0.003	0.003
Electricity			-0.485***	-0.362***	-0.434***	-0.357***	-0.357***
Eskom			0.052	-0.337	0.057	-0.096	-0.096
Agric			0.040***	0.054***	0.040***	0.050***	0.050***
Water			-0.270*	-0.433***	-0.383**	-0.478***	-0.478***
Sanitation			-0.190**	-0.235***	-0.209**	-0.255***	-0.255***
Tressindex					0.032	0.017	0.017
Distanceport							0.000
richprov							0.000
_cons	37.899**	-2.968	214.836***	207.638***	229.569***	215.963***	215.963***
Time effect	N	Y	N	Y	N	Y	Y
N	171.000	171.000	171.000	171.000	162.000	162.000	162.000
r2	0.862	0.933	0.957	0.968	0.956	0.966	0.966
r2_a	0.851	0.918	0.952	0.960	0.950	0.957	0.957
F	246.140	88.010	375.167	157.512	308.229	140.576	140.576

The dependent variable is log of provincial output (Output). Independent variables are the conventional variables which are log of labour (Emp) and log private capital (capital) and public infrastructure variables which are represented by percentage of households with tap water in yard(water); education expenditure as a percentage of provincial expenditure(edu); health expenditure as a percentage of provincial expenditure(health); agricultural loans disbursed by the Land bank(agric); percentage of household with flushed toilet(sanitation) and capital expenditure spent by Eskom the electricity utility for power mega infrastructure(eskom). Sources of data are the National Treasury for health and education data; Land bank for agriculture loans disbursed; Eskom for capex roll-out information and the rest from Quantec dataset.

For both the fixed effects model and fixed effects with instrumental variable, conventional inputs in the form of private capital and labour are positive and significant. For FE-IV, access to domestic and international markets are negative and significant. For all model, the contribution of private capital is drastically reduced when initial GDP which reflects

Table 3: Fixed Effects with IV Model

	FEIV-1	FEIV-2	FEIV-3	FEIV-4	FEIV-5	FEIV-6
Capital	0.966***	0.649***	0.655***	0.788***	0.827***	0.318***
Emp	0.022	0.062	0.044	0.211***	0.212***	0.256***
Edu	0.116	0.070	0.053	0.240***	0.297***	0.277***
Health	-0.304***	-0.052	-0.033	-0.253***	-0.293***	-0.248***
Electricity		-0.598***	-0.592***	-0.652***	-0.633***	-0.808***
Eskom		0.151*	0.111	0.341**	0.374**	0.314**
Agric			0.018	-0.058***	-0.066***	-0.036*
Water				0.571***	0.587***	0.150
Sanitation				-0.032	-0.049	0.274**
Richprov					-4.853	-1.992
Distancejhb					0.000	-0.005*
Tressindex1					-0.009	0.034*
Distanceport						-0.004
Initgdp1						6.047***
_cons	27.779*	131.807***	131.092***	6.375	3.931	-17.603
N	162.000	162.000	162.000	162.000	162.000	162.000

The dependent variable is log of provincial output (Output). Independent variables are the conventional variables which are log of labour (Emp) and log private capital (capital) and public infrastructure variables which are represented by percentage of households with tap water in yard(water); education expenditure as a percentage of provincial expenditure(edu); health expenditure as a percentage of provincial expenditure(health); agricultural loans disbursed by the Land bank(agric); percentage of household with flushed toilet(sanitation) and capital expenditure spent by Eskom the electricity utility for power mega infrastructure(eskom). The identifying instrument is the lag of private capital. FE-IV is the Fixed Effects with Instrumental Variables estimator. Sources of data are the National Treasury for health and education data; Land bank for agriculture loans disbursed; Eskom for capex roll-out information and the rest from Quantec dataset.

initial conditions is included in the model.

4.3 Dynamic Model

Lastly, we estimate the dynamic model to capture and resolve causal effects, we use Arellano and Bond (1991) difference GMM estimator. The lagged levels of the endogenous regressors are also added. This makes the endogenous variables pre-determined and, therefore, not correlated with the error term in equation. To resolve fixed effects challenge, the difference GMM uses first-differences to transform the equation. By transforming the regressors by first differencing the fixed province-specific effect is removed, because it does not vary with time. The first-differenced lagged dependent variable is also instrumented with its past levels.

The results of the dynamic model indicates that inclusion of the lagged dependent variable mops out other variables as reflected by the first model A. Capital and employment become negative and insignificant, idicating potential serial correlation and heavy trending Achen (2000). The results improve as more variables are added to the equation and capital takes on a positive sign and the lagged dependent variable is no longer too dominant as the coefficient is more than halved.

Table 4: Dynamic Model

	A	B	C	D	E
L.Output1	1.018***	1.029***	0.608***	0.551***	0.478***
L.Capital1	-0.024	-0.004	0.027	0.091	0.102
emp1	-0.033	-0.051*	-0.078***	-0.094***	-0.086***
shareedu1		-0.074	-0.060	-0.052	-0.049
sharehealth1		0.030	0.030	0.020	0.022
water1			-0.625***	-0.610***	-0.708***
electricity1			-0.117	-0.201**	-0.170*
sanitation1				0.154	0.116
agric1				0.015*	0.016*
eskom1				0.005	0.012
tressindex1					0.030*
N	153.000	153.000	153.000	153.000	153.000
F	1411.797	833.077	986.143	730.416	712.531

We used difference GMM estimator

4.4 Decomposition results

We use the regression results of equation 2 to decompose the level of inequality as measured by log variance of income. We use the results to compute the relative contribution of each explanatory variable to inequality for both the Shapley and Fields methods and compare the results.

Table 5 shows results of the two approaches suggests that the conventional production inputs, labour and private capital explain most of inequality and contribute to the worsening of regional inequality. For both methods, distance to port and rich province which to a large extent represent natural endowments contribute to inequality. Education spent by provinces appears to be an equality neutralising factor for both methods.

Results differ when it comes to Eskoms electricity infrastructure roll out, tap water in yard, sanitation and agriculural loans disbursed by the Land bank. Shapley results suggests a neutral to slightly positive contribution to inequality, whereas the Fields method suggest all of those variables assisting in reducing inequality as they have an inequality reducing effect. This results suggest that for inequality to be reduced, emphasis should be water, sanitation, agriculture and power generation and transmission. The results are similar to Zhang and Fan (2004) who showed that rural education and agricultural research and development in the western region have the most impact in reducing regional inequality in rural China.

Table 5: Decomposition Results

	Fields	Shapley
Residual	2,192	0,000
Capital	65,595	24,292
Employment	13,595	18,559
Share of prov exp - Education	0,933	0,809
Share of prov exp - Health	10,604	11,203
Electricity lighting at home	12,650	4,951
Eskom Capex rollout	-0,721	0,677
Agric - Landbank loans disbursed	-1,077	1,101
Water - tap water in yard	-6,961	2,125
Sanitation	-2,069	2,931
Rich Province	1,449	14,428
Proxy -for distance to domestic market(distancejhb)	-1,245	11,123
Proxy -for distance to international market(distanceport)	5,053	7,802
Total	100,000	100,000

Other interesting outcomes are richprovince and distance to Johannesburg as a proxy to access to domestic markets. The Shapley approach suggests that these two variables are worsening factors. This again may point to role played by initial conditions and natural endowments which tend to benefit those provinces that have access and are endowed.

4.5 Results in the context of literature

The results of this paper are largely in line with existing literature. The coefficient of initial GDP is in line with Martin and Rogers (1995). they assert that a country/region that is initially rich in capital will attract more capital. This is based on the observation that regions with better international infrastructure magnify differentials in capital endowment. In our study, inclusion of initial GDP is positive and significant and also reduces the role played by private capital. This is also similar to Costa-i Font and Rodriguez-Oreggia (2005) that found that returns to public investment differs according to regional income distribution in Mexico. The paper also found that initial GDP per capita is positive and significant for all quantiles implying that there is some path dependency in the evolution of regional income i.e. those regions that are initially rich will get more returns from public capital. These results are also consistent with Krugman (1991) who concludes and elevates initial conditions of regions as key determinant of what is likely to happen in the future.

Regarding the roll-out of electricity infrastructureDinkelman (2011) found that rural electrication has a positive effect on labour markets and that the positive impact on women is more pronounced which ties in with the decomposition results that shows that Eskoms capex roll out has an inequality reducing effect. Decomposition studies for China supports that agricultural support and water provision has an inequality reducing effect, like what the results of this study suggest.

5 Conclusion

We used a framework developed by Zhang and Fan (2004) to assess the impact of public investment on provincial inequality. We estimate a production function which includes conventional inputs and different types of public investment and use the estimated coefficient to decompose the impact of public investment on provincial inequality. The regression results indicate that rich province and initial level of GDP are both positive and significant. The positive sign of initial level of GDP is contrary to growth theory Mankiw et al. (1990) which asserts that poor regions tend to grow faster than rich regions and converge towards rich regions. In this instance, there is no evidence of convergence but rather divergence of regions.

The estimated coefficients are used to decompose impact of public investment on provincial inequality using both the Fields and Shorrock methods. The decomposition results suggests that water, sanitation, agriculture and Eskom's capex roll out are the most efficient public investment in the fight against provincial inequality. Initial conditions and geography as represented by initial GDP and distance to domestic markets favour those that are better endowed and contribute to the worsening of inequality.

References

Achen, C. H. (2000), 'Why lagged dependent variables can suppress the explanatory power of other independent variables', *Ann Arbor* **1001**(2000), 48106–1248.

Arellano, M. and Bond, S. (1991), 'Some tests of specification for panel data: Monte carlo evidence and an application to employment equations', *The review of economic studies* **58**(2), 277–297.

Aschauer, D. A. (1989), 'Is public expenditure productive?', *Journal of monetary economics* **23**(2), 177–200.

Baltagi, B. (2008), *Econometric analysis of panel data*, John Wiley & Sons.

Barro, R. J. (2004), 'Sala-i-martin x (2004) economic growth', *Second Editions, Mc-Grow-Hill, New York* .

Chatterjee, S. and Turnovsky, S. J. (2012), 'Infrastructure and inequality', *European Economic Review* **56**(8), 1730–1745.

Costa-i Font, J. and Rodriguez-Oreggia, E. (2005), 'Is the impact of public investment neutral across the regional income distribution? evidence from mexico', *Economic Geography* **81**(3), 305–322.

Demurger, S. (2001), 'Infrastructure development and economic growth: an explanation for regional disparities in china?', *Journal of Comparative economics* **29**(1), 95–117.

Dinkelman, T. (2011), 'The effects of rural electrification on employment: New evidence from south africa', *The American Economic Review* **101**(7), 3078–3108.

Easterly, W. and Rebelo, S. (1993), 'Fiscal policy and economic growth', *Journal of monetary economics* **32**(3), 417–458.

Fedderke, J. W., Perkins, P. and Luiz, J. M. (2006), 'Infrastructural investment in long-run economic growth: South africa 1875–2001', *World development* **34**(6), 1037–1059.

Fields, G. S. (2003), 'Accounting for income inequality and its change: A new method, with application to the distribution of earnings in the united states', *Research in labor economics* **22**(3), 1–38.

Getachew, Y. Y. (2010), 'Public capital and distributional dynamics in a two-sector growth model', *Journal of Macroeconomics* **32**(2), 606–616.

Getachew, Y. Y. and Turnovsky, S. J. (2015), 'Productive government spending and its consequences for the growth–inequality tradeoff', *Research in Economics* **69**(4), 621–640.

Heshmati, A. (2004), 'A review of decomposition of income inequality', *IZA Discussion Paper* .

Klenert, D., Mattauch, L., Edenhofer, O. and Lessmann, K. (2014), 'Infrastructure and inequality: insights from incorporating key economic facts about household heterogeneity'.

Krugman, P. (1990), Increasing returns and economic geography, Technical report, National Bureau of Economic Research.

Krugman, P. (1991), 'History and industry location: the case of the manufacturing belt', *The American Economic Review* **81**(2), 80–83.

Mankiw, N. G., Romer, D. and Weil, D. N. (1990), A contribution to the empirics of economic growth, Technical report, National Bureau of Economic Research.

Manna, R., Regoli, A. et al. (2012), 'Regression-based approaches for the decomposition of income inequality in italy, 1998–2008', *Rivista di Statistica ufficiale* (1).

Martin, P. (1999), 'Public policies, regional inequalities and growth', *Journal of public economics* **73**(1), 85–105.

Martin, P. and Rogers, C. A. (1995), 'Industrial location and public infrastructure', *Journal of international Economics* **39**(3), 335–351.

Morduch, J. and Sicular, T. (2002), 'Rethinking inequality decomposition, with evidence from rural china', *The Economic Journal* **112**(476), 93–106.

Perkins, P., Fedderke, J. and Luiz, J. (2005), 'An analysis of economic infrastructure investment in south africa', *South African Journal of Economics* **73**(2), 211–228.

Phillips, P. C. and Moon, H. R. (1999), 'Linear regression limit theory for nonstationary panel data', *Econometrica* **67**(5), 1057–1111.

Roodman, D. (2006), 'How to do xtabond2: An introduction to difference and system gmm in stata'.

Shorrocks, A. F. (1982), 'Inequality decomposition by factor components', *Econometrica: Journal of the Econometric Society* pp. 193–211.

Shorrocks, A. F. (1984), 'Inequality decomposition by population subgroups', *Econometrica: Journal of the Econometric Society* pp. 1369–1385.

Shorrocks, A. F. (2013), 'Decomposition procedures for distributional analysis: a unified framework based on the shapley value', *Journal of Economic Inequality* pp. 1–28.

Wan, G. and Zhou, Z. (2005), 'Income inequality in rural china: Regression-based decomposition using household data', *Review of development economics* **9**(1), 107–120.

Zhang, X. and Fan, S. (2004), 'Public investment and regional inequality in rural china', *Agricultural Economics* **30**(2), 89–100.