UNIVERSIDAD PONTIFICIA COMILLAS

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

OFFICIAL MASTER'S DEGREE IN THE
ELECTRIC POWER INDUSTRY

Master’s Thesis

Economic and Social Effects of Infrastructure-driven Development in Africa

Author: **Philippe Peelman**

Supervisor: **Guillaume de Chorivit**

Madrid, June 2016

Master's Thesis Presentation Authorization

THE STUDENT:

Philippe Peelman

....

THE SUPERVISOR

Guillaume de Chorivit

Signed: Date: 27/ 06/ 2016

THE CO-SUPERVISOR

NAME OF THE CO-SUPERVISOR, or leave it in blank

Signed: .... Date: ....../ ....../ ......

Authorization of the Master's Thesis Coordinator

Dr. Luis Olmos Camacho

Signed.: .... Date: ....../ ....../ ......

UNIVERSIDAD PONTIFICIA COMILLAS

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

OFFICIAL MASTER'S DEGREE IN THE
ELECTRIC POWER INDUSTRY

Master's Thesis

Economic and Social Effects of Infrastructure-driven Development in Africa

Author: **Philippe Peelman**

Supervisor: **Guillaume de Chorivit**

Madrid, June 2016

Summary

This thesis attempts to estimate the impact of infrastructural developments on people's livelihoods. Two different infrastructural investment areas and two different interpretations of livelihood improvements are examined.

Firstly, two distinct infrastructure types are being examined: telecommunications and electricity generation/consumption. Both are abundantly discussed in the literature but there is still no consensus on the exact impact of its effect on GDP and other developmental indices. Secondly, two different livelihood factors are analyzed: GDP per capita and Female Genital Cutting (FGC). GDP per capita is a pure income measure while FGC is the result of complex social and historical prerogatives.

Additionally, the effects between infrastructure and livelihood are being researched from two perspectives: a macro (data on country-level) and micro (data on the level of the individual) view.

The whole thesis and subsequent analysis is centered around Africa.

At the macro level, the example of South Africa was used to detect the direction of causality between GDP per capita and respectively electricity consumption and mobile penetration. A modified Sims test was applied to the data but no unilateral or bilateral causality could be detected.

At the micro level, techniques of machine learning were applied to the USAID and UNICEF data for a large set of African countries. While a lot of determinants were identified, the hypothesis that increased mobile penetration could have a measurable impact of the abandonment of FGC, could not be confirmed.

Table of Contents

List of Figures

1. Introduction

In the past decade, Africa, and especially sub-Saharan Africa, has undergone an impressive mobile revolution. While the continent always lagged behind in the necessary telecommunications investments needed for landlines to reach a critical mass, it has recently become the second largerst mobile phone market in the world.

Mobile telecommunications has transformed the continent. Besides the typical use-cases, its influence has also been felt in a number of seemingly unrelated areas. For example, mobile money, pioneered by Safaricom in Kenya and adopted by the majority of telecom providers in sub-saharan Africa, has provided financial services to the single largest number of unbanked individuals in this world.

Traditionally, the advent of mobile telephony in Africa was discussed from a purely economic perspective, where it fits in in the decades-old infrastructure debate: large social overhead investments undertaken by the government are hypothesized to be strongly driving economic prosperity. However, an important aspect makes telecommunications a rather interesting case: telecommunications infrastructure becomes increasingly valuable with the number of users: this phenomenon is called a direct network externality (**Economides and Himmelberg, 2005).**

This thesis also investigates the possibility that increased telecommunications penetration might have benefits other than strictly economical. Female genital cutting/mutiliation is a deplorable practice which is the result of deeply rooted societal beliefs and has survived the generations. The result of this practice, however, is something which is clearly testable and non-ambiguous. It will therefore be hypothesized later on in this document that the availability of outside ideas in a close-knot society might have a profound influence on the decline of the practice.

2. Literature

2.1. Infrastructure Investment

Since the '90s, strong evidence has emerged finding that the existence of policies which promote capital accumulation and the rate of investment of an economy, is associated with future economic growth rates (**Delong and Summers (1991), Levine and Renelt (1992) and Fischer et al. (1996a), (1996b)**). Furthermore, it has been argued by **Aschauer (1989)** and **DeLong and Summers (1991)** that certain types of investment, in particular public infrastructure investments are more strongly associated with productivity and growth. When using the term 'infrastructure' in economic research, one generally refers to *large social overhead capital* (**Nandi and Nadiri, 2003**). This includes but is not limited to roads, electricity generation, telecommunication networks and sewers. Investments in this area are often characterized by their long-term and uncertain nature, causing them in most cases to be undertaken by the public sector.

Since the start of the infrastructure-output research area in the first part of the 1960s, the exact direction of causality has been fiercely debated. There are three competing theoretical frameworks which explain the direction of causality between investment and output. Firstly, an aggregate production function can be made to model output as a function of capital, labor and technology. Investment in infrastructure is perceived as an increase in capital which unilaterally increases output. This is known as the 'growth hypothesis'. A classic production function model was proposed by **Ram (1986).** However, the **acceleration principle** or conservation hypothesis, as most recently represented in **Cullem (1998)**, hypothesizes that aggregate income is the main driver of investment, thus reversing the direction of causality. Lastly, models as in **Roller and Waverman (1996)** propose a system of bilateral causality, where feedback loops exist between both output and investment (feedback hypothesis). Nowadays, it is generally hypothesized that two directions of causality often exist simultaneously for most infrastructure investments:

- An increase in infrastructure investment triggers economic growth through its **direct** and **indirect** effects (as discussed infra in the case of telecommunications investment)
- Higher economic output increases the demand for infrastructure (eg. **income elasticity** of transport/mobile services/etc.)

In order to make the discussion more concrete, two main modes of infrastructure investment will be discussed in the following sections: **telecommunications** and energy (**electricity**).

2.2. Telecommunication

There exist a limited number of studies which attempt to gauge the income impact of telecommunications infrastructure development at the micro-economic level. For example, T**olero et al. (2003)** found that telecommunication projects in Bangladesh and Peru are generally welfare enhancing for rural households. They determined this by measuring the willingness to pay which is found to be higher than the prevailing tariffs. In the wake of privatization, **Chong et al. (2006)** took advantage of the random installation of public telephones in rural villages in Peru to conduct a natural experiment. They found that public telephone usage is linked with income measurements.

However, while their research design is of great value, these small-scale studies are limited by their scope in time and space. As a compromise, large econometric studies are daily being undertaken to determine the direction of causality and the magnitude of the effect for different contexts and types of telecommunications investments. A separate section in this thesis is dedicated to this kind of studies and the econometrics employed in them.

Investment in telecommunication aligns with other infrastructure investments by means of the **direct** effect it has on GDP. The establishment of, for example, a fiber grid uses raw materials and man-hours which increases demand in telecommunication and related industries, while also promoting growth through the economic multiplier. Furthermore, infrastructure investments have a profound influence on economic growth by increasing the productivity of unrelated businesses and agents through so-called **indirect** effects. It is here where in the large field of infrastructure research (transportation, electricity etc.), telecommunication clearly stands apart from the others and has its greatest effect on economic growth. This is because of the existence of a ***direct network externality*** (**Economides and Himmelberg, 2005)**. This positive externality entails that the value of the network increases with its user base and as a result amplifies any indirect effects when increasing telecommunication investment.

2.2.1. Indirect effects

The exact mechanisms by which these indirect effects manifest themselves are debated in numerous theoretical papers but unfortunately far less often validated by means of an empirical study.

It has been established that an increase in the ease and affordability of telecommunication promotes market efficiency by reducing price dispersion among markets. In the absence of collusion, price differentials have their cause in the existence of transaction costs. A **transaction cost** is '*the cost inherent in making an economic exchange'*, and has **search costs** as one of its components. Buyers (resp. sellers) are continuously searching to maximize their consumer (resp. producer) surplus but in doing so they have to incur costs in order to acquire the relevant price information (search costs). **Transportation costs** are closely linked to search costs because in a pre-telecommunication society transportation is essential to the exchange of information. As **North, 1995** remarked: '*information is costly and asymmetrically held by the parties of exchange'.* This phenomenon where transaction costs (more specifically search costs) and asymmetric information causes price dispersion for homogenous goods was first introduced by **Stigler (1961)** in his work 'Economics of Information'. Homogenizing prices has multiple beneficial results: increasing producer and consumer surplus (as discussed supra), arriving at a smoother price pattern year-round, establishing a stable supply of products and services, less spoilage of perishables etc.

Garbade and Silber (1978) found statistical evidence that inter-market price differentials decreased after two large innovations in ICT (Telegraph and Trans-Atlantic cable). **Bayes (2001)** researched the effect of the leasing of phones in a rural context in Bangladesh. He concluded that it greatly enabled the decrease of transaction costs, and that the initiative disproportionally benefitted the poor, as measured by their consumer surplus (defined in the study as *'the price users would have paid for an alternative mode of communication, minus the price they actually paid for the phone service'*). **Overa (2006)** conducted numerous interviews over a period of 3 years to examine the changing trading practices in Ghana due to the wide introduction of telecommunications. All traders found the exchange of information on prices, supply and demand to be the most important utility of mobile communication. Secondly, mobile phones are being used to improve the coordination of multi-local business activities (eg. onion trade in Ghana). Finally, he found that the risk of money transactions could be reduced in Ghana's frozen fish industry by communicating instantly with mediating financial institutions. **Jensen (2007)** found that the wide availability of inexpensive mobile telephony caused price differentials to decrease on local fish markets in South India. **Aker (2008)** found evidence that the introduction of mobile phones decreased price dispersion on grain markets in Niger with minimum 6%. Price dispersion decreased more for markets which were farther away or for which the road quality was worse. In Zinder, Niger a grain trader illustrated the influence of search costs in his business in the following way: "*(With a cell phone), I know the price for $2, rather than travelling (to the market), which costs $20*."

Muto (2008) showed that mobile phone coverage can promote market participation of farmers in remote areas who produce perishable goods. He found an increase of 50 to 69 percent of market participation for banana farmers who lived more than 20 miles away from district centers.

Matambalya et al. (2001) analyzed the effect of information technologies on the performance of small and medium enterprises and suggest that there may be no effect. On the other hand, **Leff**

(1984) asserted that improved telecommunication allows firms to have more geographically dispersed activities which provides them with the benefit of establishing economies of scale and scope.

At a more socio-economic level, **Bayes (2001)** also provided evidence that phones leased to women in Bangladesh might help with their empowerment in the society. For the society at large, law enforcement became more effective (eg. victims of burglary in remote villages were able to contact the police more rapidly) and communication during disasters was improved (as observed during the flood in July-September 1998).

2.2.2. Livelihood Framework

While most quantitative research points to a large and substantial growth effect of telecommunication investment (as will be reviewed infra), there is a profound lack of any sort of structural framework of the indirect effects of the resulting infrastructure. As **Heeks (2007)** put it: *"there has been a bias to action, not a bias to knowledge. We are changing the world without interpreting or understanding it"*. **Duncombe (2012)** attempted to provide such a framework using elements from **livelihood analysis** and focusing on mobile telephony in developing countries. In livelihood analysis, assets are found to play a vital role in extending the decision space available to people in vulnerable contexts (**Ellis & Bahiigwa, 2003; Ellis, 2000).** Asset status is impacted by mobile phones in 4 main ways:

1. **Assets substitution**

   Improved long-distance telecommunication has caused the substitution between transportation costs and mobile phone airtime. A cattle herder, for example, no longer needs to leave his herd when he/she wants to report a sick animal. **Bayes (2001)** reported that when asked "*how would you satisfy the need for current calls had there been no VPP [*mobile phone*] in your village*?", 68% of respondents from the poorest population segment in Bangladesh said to opt for physical mobility instead. This has as a result that, despite the (still) high cost of mobile phone technology, women, for example, are substituting normal purchases with mobile phones and credit, in order to pursue a **long-term asset accumulation strategy**.

2. **Asset enhancement**

   Mobile phones permit the use of scarce resources with greater efficiency. In organizationally complex agricultural value chains, for example, **information search** and **coordination** of geographically dispersed activities is greatly enhanced.

3. **Asset disembodiment**

   Crucial contact information changes its nominal form and becomes embedded in the new mobile technology. After having previously held in memory or recorded informally in a notebook, it becomes codified, recorded, formalized and even shared. It helps to solidify existing **social capital resources**.

4. **Asset exchange and combination**

M-Pesa, Safaricom's service in Kenya, spearheaded the boom of mobile financial services (MFS) on the African continent. MFS (also aptly named 'mobile money') allows the use of mobile airtime as a cash substitute. As a result, mobile technology created and facilitated a new means of **financial exchange.**

2.3. Energy (Electricity)

"Energy access" is commonly defined as the possibility to secure modern (as consumed in developed countries) forms of energy at affordable prices (**Bhattacharyya, 2006, in press**; **Spalding-Fecher et al., 2005**). This is of vital importance, because "*a country's economy and its energy use, particularly electricity use, are linked. Short-term changes in electricity use are often positively correlated with changes in economic output"* (**EIA, 2013b).**

This idea of a correlation between economic prosperity and shifting energy use is embodied within the **energy ladder**. An energy ladder typically shows the correlation between modes of energy and prosperity.

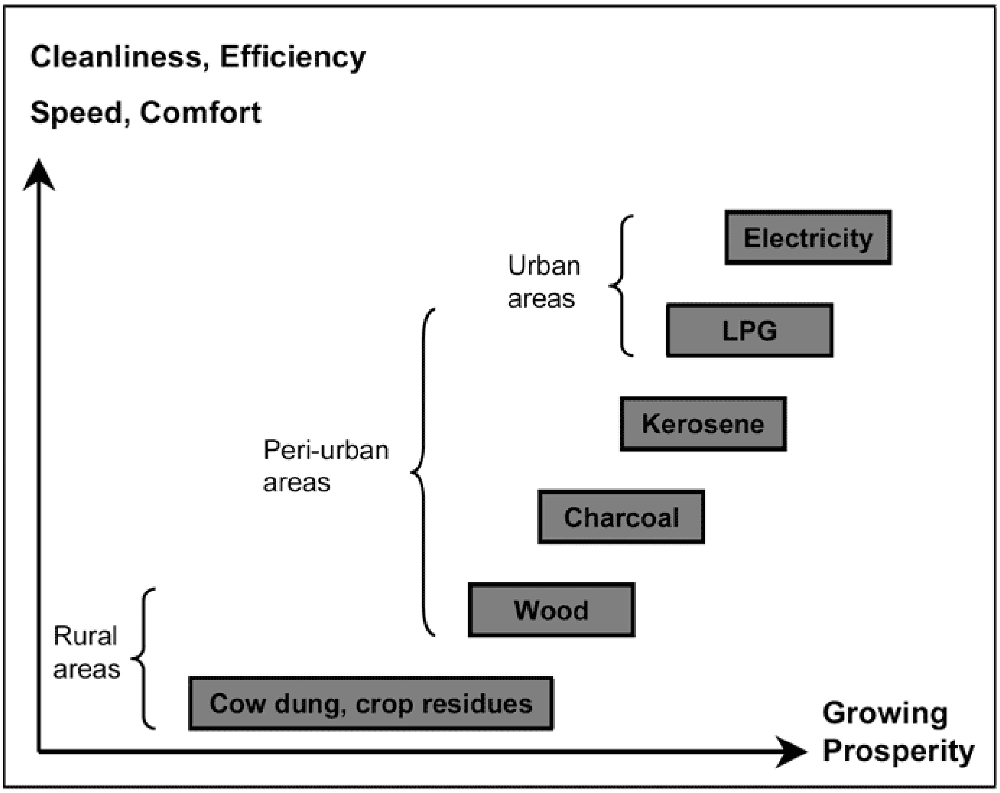


Figure 1: The Energy Ladder

The poorest of households typically spend a lot of time daily in order to gather firewood or use other low energy-dense forms of biomass. These forms of energy often damage the household members' health after long-term improper use. More affluent households are able to buy charcoal or oil-based derivatives. This is still a decentralized process of energy conversion with each household procuring their own supply of fossil fuels and converting it to heat/light independently. At the top of the ladder we encounter electricity. Infrastructure constraints make this mode of energy generally appear first in urban zones. It is clean, efficient and flexible.

It is generaly accepted that the use of more modern forms of energy (electricity) in increased volumes brings multi-dimensional benefits**.** In **Kanagawa and Nakata (2008)**, 4 dimensions are enumerated:

Health

Burning of biomass in enclosed spaces in less than efficient stoves produces polluants, as for example carbon monoxide. A particular interesting study (**WIN News, 1998**) in The Gambia found that girls below the age of five which are routinely carried by their mothers while cooking have a 6 time higher risk of lung cancer than if this was not the case and their parents smoked instead.

Education

Current research has established that inequality of income negatively affects opportunities of education. This is because poor households are less equipped to carry direct and indirect educational expenses. This ultimately leads to low-income families being gripped by a vicious circle of poverty. This is of particular societal concern because education has been identified as a **positive externality**: while costs are typically envisaged by the school-goer, benefits of an educated populace accrues to the entire society. Electricity access has a substantial effect on education. Children are liberated from part of the housework and can expand their opportunities for school attendance. Furthermore, electrification and modern lighting enables night-time study and the use of TV, radio and internet.

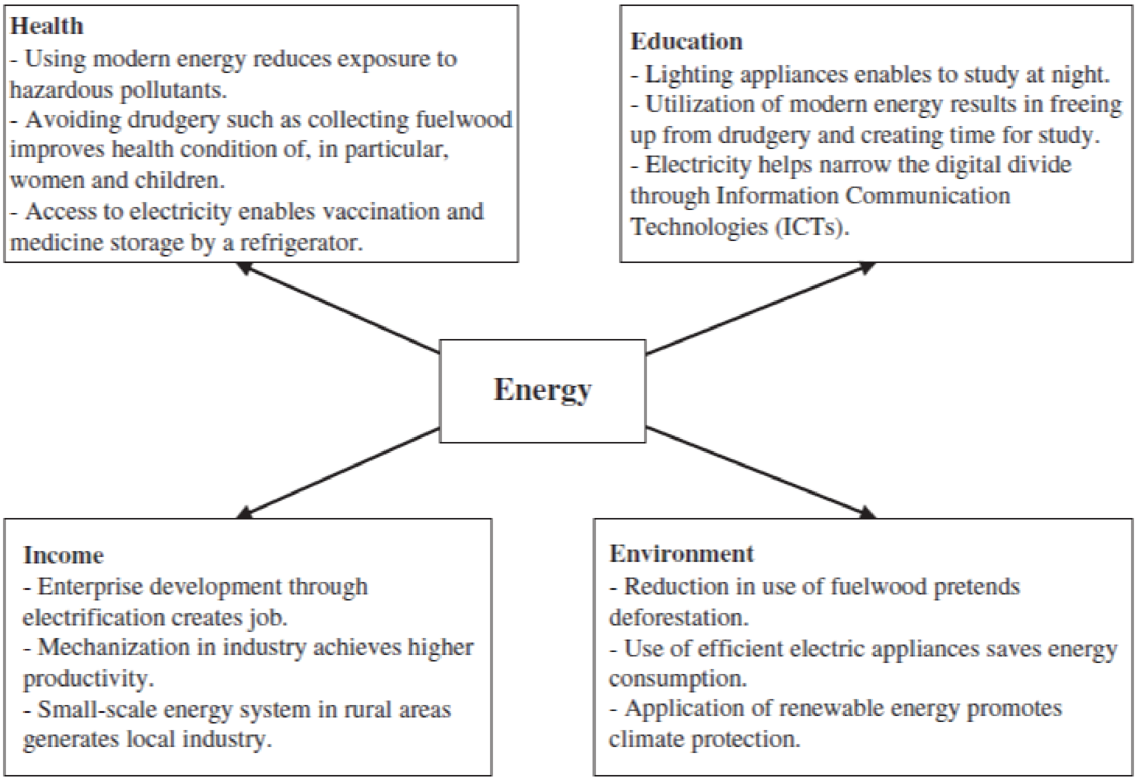
Environment

In a first step, centralized electricity production increases upon the efficiency of individualized, traditional combustion processes of firewood, coal etc. This is combined with a decrease in the green house gasses (GHG) produced per kWh. A society depend on the burning of biomass is also linked to problems of deforestation and desertification.

Income

Access to electricity frees up women and girls to engage in other, more productive activities. This is accomplished by eliminating certain activities (gathering of firewood) and drastically reducing certain housework tasks (laundry, cleaning etc.). Women and girls can substitute that time by taking part in the labor market, establishing small enterprises or attend schools (or other educational activities). This has a profound effect on income, as well on empowerment of women (which are typically responsible for housework in African societies and the world) **DfID (2002)**.

As can be seen from the below figure, **urbanization** is typically identified as an instrumental variable in this relationship. **Niu et al. (2013**) and **Mihra et al. (2009)** found a feedback relation between electricity consumption and urbanization. On the other hand, for China, **Liu (2009)** found only unidirectional Granger causality running from urbanization to electricity consumption in the short and long term. However, it has to be said that most research has only focused on the assumption that urbanization could potentially increase GHG emissions through rapidly expanding fossil fuel use.



Health
- Using modern energy reduces exposure to hazardous pollutants.
- Avoiding drudgery such as collecting fuelwood improves health condition of, in particular, women and children.
- Access to electricity enables vaccination and medicine storage by a refrigerator.
Education
- Lighting appliances enables to study at night.
- Utilization of modern energy results in freeing up from drudgery and creating time for study.
- Electricity helps narrow the digital divide through Information Communication Technologies (ICTs).
Energy
Income
- Enterprise development through electrification creates job.
- Mechanization in industry achieves higher productivity.
- Small-scale energy system in rural areas generates local industry.
Environment
- Reduction in use of fuelwood pretends deforestation.
- Use of efficient electric appliances saves energy consumption.
- Application of renewable energy promotes climate protection.

Figure 2: Multi-dimensional benefits of modern modes of energy

From a theoretical viewpoint, it is clear that increased electricity access and consumption may have an effect on economic prosperity. A large number of studies has attempted to unveil this causal relationship on the macro level:

Wolde-Rufael (2004) investigated the long-run and causal relationship between electricity use per capita and GDP per capita for 17 African countries between 1971 and 2000. They found a positive unidirectional causality running from electricity use to GDP for 6 countries, an opposite causality for 3 countries and a bidirectional causality for another 3.

Karanfil and Li (2014) researched the short- and long-term relationship between electricity consumption and economic activities for 160 countries between 1980-2010. They used a multivariate framework including net import of electricity (as proxy for electricity dependence) and urbanization. They found a long-run co-integration relationship (feedback hypothesis) between electricity consumption and economic growth in the full sample and majority of subsamples. In the short term, they found evidence of unidirectional causality from economic growth to electricity consumption in most lower-income panels and no causality in the higher-income panels.

To illustrate the diverse results obtained in a wide variety of studies examining the causal relationship between electricity consumption and economic growth, a short list of often-quoted studies is presented below:

Year	Author	Region	Result
1997	Glauser and Lia	South Korea and Singapore	Bidirectional
2000	Yang	Taiwan	Bidirectional

2002	Gosh	India	Unidirectional: econ. growth to electr. cons.
2004	Morimoto and Hope	Sri Lanka	Bidirectional
2004	Jumbe	Malawi	Bidirectional
2004	Fatai et al.	Australia	Unidirectional: econ. growth to electr. cons.
2004	Thoma	USA	Unidirectional: econ. growth to electr. cons.
2004	Shiu and Lam	China	Unidirectional: electr cons. to econ. growth
2004	Wolde-Rufael	Sjanghai	Unidirectional: electr cons. to econ. growth
2005	Narayan and Smyth	Australia	Unidirectional: econ. growth to electr. cons.

2.4. Female Genital Cutting (FGC)

Female Genital Cutting (FGC) includes procedures that intentionally alter or cause injury to the female genital organs for non-medical reasons **(WHO, 2014)** and has been deemed a violation of human rights. It is an old practice, referred to in pharaonic writings. (**Snow et al., 2002)** The practice differs in incidence by region and ethnicity, age at circumcision, the type of practitioner who performed the circumcision and affiliation to religious organizations. There is a large incidence of the practice in numerous countries in Western and Northeastern Africa.

2.4.1. Terminology

WHO (2008) overviewed all terminology associated with the circumcision act. When the practice was first observed by outsiders, the term '**Female Circumcision (FC)'** was used. However, because male circumcision is an altogether distinct practice, there was concern that the term would cause confusion about the severity and acceptability of the procedure. The term '**Female Genital Mutilation (FGM)'** made its introduction in the late 1990s and was deemed to have three main advantages over the previous name. Firstly, it allows to clearly distinguish from male circumcision. Secondly, the 'mutilation' part stresses the invasiveness and severity of the procedure. Thirdly, it refers to the practice's violation of human rights and tries to generate dynamic action against the practice. Nonetheless, multiple African experts rejected the addition of 'mutilation' to the term as it implied deliberate harm. They suggested the term **Female Genital Cutting (FGC)**, as more neutral. As a compromise, some agencies (among which UNICEF) began using a composite of FGM and FGC: **'Female Genital Mutilation/Cutting (FGM/C)'**. In this thesis, the term **Female Genital Cutting (FGC)**, as used by the **DHS**, will be utilized.

2.4.2. Types of FGC

WHO (1997) distinguishes four types of FGC:

1) **Type 1**: *'the partial or total removal of the clitoris and/or the prepuce (clitoridectomy)'*
2) **Type 2**: *'the partial or total removal of the clitoris and the labia minora, with or without excision of the labia majora (excision)'*
3) **Type 3:** *'narrowing of the vaginial orifice with creation of a covering seal by cutting and appositioning the labia minora and/or the labia majora, with or without excision of the clitoris (infibulations)'*

4) **Type 4**: *'all other harmful procedures to the female genitalia for non-medical purpose (other)'*

Type 3, which involves infibulation (the sewing or sealing (nearly) shut of the vaginal orifice), is of most concern, for its great potential to harm the health of the individual undergoing the procedure (**Obermeyer, 2005**). Women who are infibulated frequently experience pain and damage to their genitalia during pregnancy and delivery.

In the DHS surveys, three simplified categories of circumcisions are utilized: (1) Nicked only, (2) Flesh removed or (3) Sewn closed.

2.4.3. Short history

Year	Event
1979	**Francis Hosken** provides the first estimate for the national prevalence of FGC in many African countries
1989-1990	The **DHS** survey in northern Sudan contains the first FGC questions posed to female respondents
1997	The first national figures on FGC prevalence in six African countries (with Yemen) is published by **Carr**
2005	The Multiple Cluster Indicator Surveys (MICS) conducted by UNICEF begins including a module on female genital cutting in selected countries
2010	DHS and UNICEF coordinate their respective FGC modules and begin gathering data for all living daughters

2.4.4. Data Sources

USAID, through its **Demographic and Health Surveys (DHS)** and UNICEF via its **Multiple Cluster Indicator Surveys (MICS)** are the main international organizations involved in data collection regarding FGC. DHS first included a FGC module in 1989-1990, while MICS only started asking questions regarding FGC in 2005. Since 2010, data on all living daughters is being collected by MICS and DHS. The proper use of this data is often called an '*enduring challenge*' by DHS and MICS (**DHS, 2013).**

2.4.5. Validity of self-reporting

Both sources of FGC data (DHS and MICS) rely on **self-reported data** opposed to **clinical examination**. The assumptions on which such surveys rely on are that interviewed women know whether they have undergone (a specific type of) circumcision and if they will truthfully and accurately share that information with the interviewer. Over the past 20 years, various studies have verified self-reported incidence rates by means of physical examinations or follow-up interviews. Results vary wildly.

Publication year	Author(s)	Location	Accuracy of FCG statement (%)
1989	Odujinrin et al.	Nigeria	75% of circumcised women
1996	EFCS	Egypt	94% of interviewed women
1997	Adinma	South-East Nigeria	57% of interviewed women
2003	Jackson et al.	Northern Ghana (Kassena-Nankana district)	85% of interviewed women

In most cases, an answer category of 'Not sure' is also provided. With regard to this, **Odujinrin et al. (1989)** found that the vast majority (81%) of 'unsure' women were uncut.

Self-reporting validity is also greatly affected by the existence or emergence of a law against circumcision. Women will less readily admit to have been circumcised after the law is in effect. The impact of such a law can be estimated by comparing FGC incidence for age cohorts before and after passing of the law. The effect of a law on respondents' answers depends on (1) when the law was passed, (2) how widely known the law is, (3) the willingness of the state to prosecute and (4) the expectations of citizens concerning possible punishment.

2.4.6. Determinants

The most significant social predictor of FGC is **ethnicity**. In countries with a national prevalence of less than 80%, large differences can be found between ethnical groups.

In most countries, the practice is more prevalent in **rural areas** compared to urban areas (Senegal, Burkina Faso, Ethiopia, Kenya, Tanzania). Nigeria is an outlier in this respect because of the fact that some rural ethnical groups do not practice FGC, skewing the results. A likely theory into the cause of this phenomenon could be that families in urban areas are less likely to follow the requests of older family members who would normally guide the circumcision rituals (no study has yet been conducted to empirically verify this).

Also, daughters of mother with higher levels of **education** are less likely to undergo FGC. This is possibly also true for **income.** Before DHS asked respondents' about the FGC status of their daughters, the correlation between FGC status and education/income of the mother was routinely tabulated and published. This practice was discontinued due to the fact that the FGC act clearly predates the educational and financial attainment of the mother (**DHS, 2013)**. On the other hand, a possible explanation for the link could be that could be that girls who are circumcised grow up in more conservative households and receive less opportunities regarding receiving an education.

While there is some confusion among the general populace about the religious aspect of the act, **Obermeyer (1999)** found the role of religion as a predictor *'not at all consistent'* and '*without any theological foundation'*. In some cases, however, religion has been used as a rationale for performing FGC. Because of this phenomenon, continuous efforts have been made to persuade religious leaders to actively promote **against** the practice.

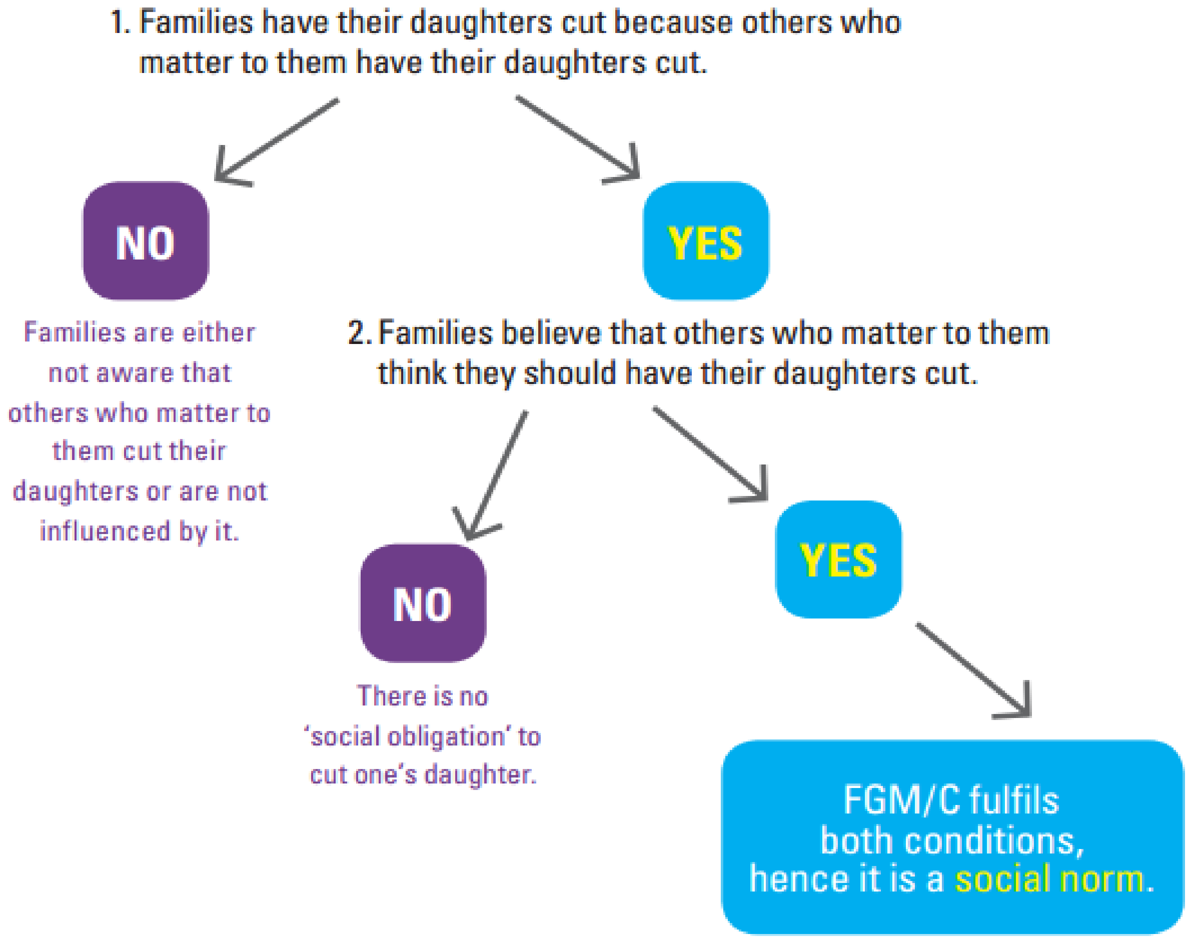
A root, psychological cause into the continuation of the practice to this day can be found in the concept of **social acceptance**. Questions related to the benefits of FGC in DHS' FGC module, reveal that **social acceptance** is often considered the greatest benefit of undergoing the practice.

2.4.7. Telecommunication as a Driver

Jensen and Oster (2009) found that the spread of cable and satellite television in India had siginificant effects on the empowerment of women in India. They used a 3 year individual-level panel data set at the time of a staged roll-out of cable television in rural parts of India. They found that when women were exposed to new information about the outside world and other ways of life, they were less tolerant of domestic violence against women, mothered less children and were less inclined to have a strong son preference.

With the recent mobile telephony revolution and emergence of low-cost internet bundles on the African continent, the mobile phone has become the number one conveyer of external information to a large number of rural women. The mobile phone could be, as it were, the equivalent of cable television in India.

Furthermore, the MICS found that the practice of FGC was largely held in place by conservative **societal norms** and the resulting **peer pressure**:



1. Families have their daughters cut because others who matter to them have their daughters cut.

Families are either not aware that others who matter to them cut their daughters or are not influenced by it.

2. Families believe that others who matter to them think they should have their daughters cut.

There is no 'social obligation' to cut one's daughter.

FGM/C fulfils both conditions, hence it is a social norm.

Just as Indian women become aware that female empowerment has more or less become the norm in Western society and as a result no longer accepted to be treated badly by their husbands, African women possessing a mobile phone could be given an outside view on this outdated practice and abandon it altogether.

This is one of the **hypotheses** which will studied in this thesis.

2.5. Discussion of Econometric Techniques

A lot of studies have correlated telecommunications (typically telephones per 100 people) with GNP per capita (**Jipp, 1932**; **Hardy, 1980**). Notably, in 1963, Jipp published an article in the International Telecommunication Union journal entitled "Wealth of nations and telephone density". In the article, he compared for the first time the ratio of teledensity to per capita income levels. He found a large positive correlation between per capita income and teledensity. This relationship was later dubbed "Jipp's law". Jipp intended the found relationship to be leveraged as a tool to determine the optimal telecommunication investment in the then state-owned monopoly telecommunications firms.

Traditionally (and also used in this paper), a strong linear relationship is obtained when logarithmically transforming one or both of the time series. This implies that the relationship is probably nonlinear and of the functional form $y = a*x^b$. This corresponds with the conventional functional form of a production function (eg. Cobb-Douglas). As such, it is entirely possible that the relationship has other characteristics of traditional production functions, for example diminishing returns.

Using data of the International Telecommunication Union (ITU), this correlation is pictured in figure 3 and figure 4. In both figures, the X-axis depicts the natural logarithm of GDP per capita while the Y-axis depicts Mobile penetration per 100. In figure 3, each data point represents one of 207 countries for which data was available. The blue data points correspond with the year 2000 while the red data points correspond with the year 2013. It can be seen that there exists a clear positive relationship between all countries for each of the two years. This relationship is more pronounced for the year 2013, because of the longer timeframe since worldwide introduction of the new mobile technology.

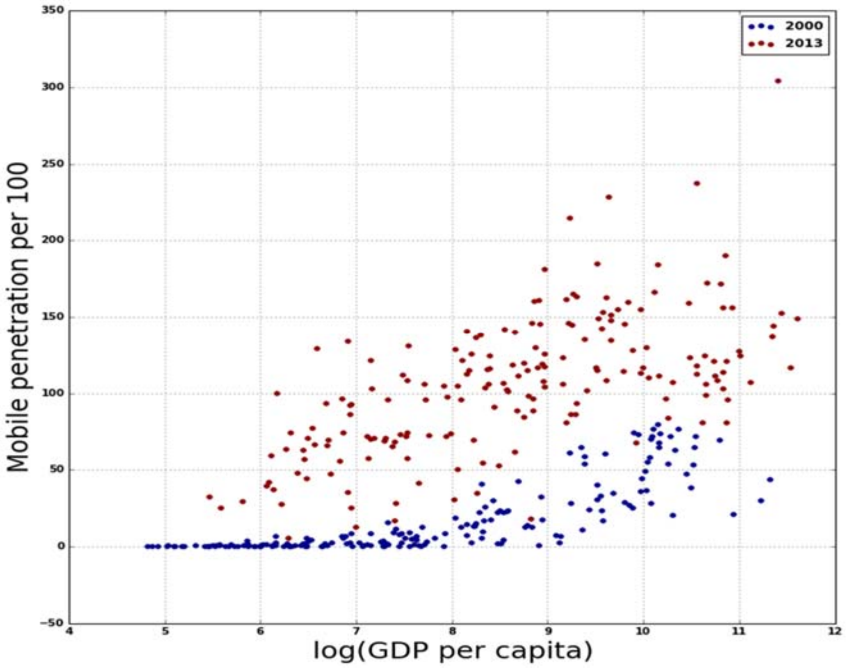


Figure 3: Jipp's law for 207 countries (2000 and 2013)

Figure 4 looks at the correlation within each country. In order to not overcrowd the diagram, only four countries are depicted: Spain, Rwanda, Egypt and Rwanda. As can be seen, this positive relationship is even more pronounced within each country.

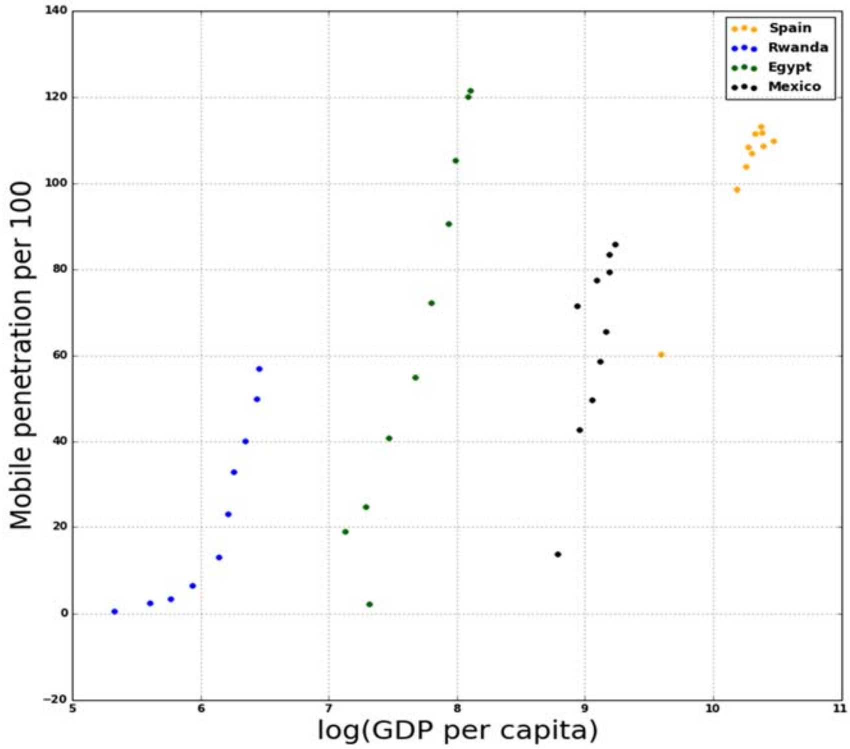


Figure 4: Jipp's Law for Spain, Rwanda, Egypt and Mexico

Ofcourse, correlation does not imply causation. Especially the theorized two-directional interrelationship between telecommunication investment and income complicates any econometric analysis.

2.5.1. Multiple regression analysis (single equation model)

Multiple regression analysis is a statistical technique to model the relationship between one dependent and multiple (>2) independent variables. Mathematically:

$$\boldsymbol{Y_{it} = \beta_1 X_{it1} + \beta_2 X_{it2} + .. + \beta_p X_{itp} + u_{it}}$$

Publication Year	Author(s)	☎	📱	Region	Years	Result(s)
1968	CCITT group	X		World	1955,1960, 1965	Significant, positive: elasticity of 1.4
1976	Bebee and Gilling	X		29 countries	1970	Significant and positive
1980	Hardy	X		15 developed and 45 developing countries	1960-1973	Significant, positive, but not significant for business telephones and developed/developing separately
1992	Norton	X		47 countries (developed and developing)	1957-1977	Significant, positive
1994	Dhokalia and Harlam	X		50 states of the USA	1985, 1990	Significant, positive
1998	Madden and			27 CEE countries	1991-1994	Significant, positive; 1% TEL results in

	Savage					14% of growth. Two-way causality
2004	Datta and Agarwal	X		22 OECD countries	1980-1992	Significant, positive. Unilateral causality, diminishing returns

The CCITT group (1968) modeled a **log linear relationship** between the telephone line density and GDP per capita. Data of the ITU (International Telecmmunications union) for years 1955, 1960 and 1965 was used. Their analysis showed that the elasticity of teledensity vis-à-vis per capita GDP is greater than unity (1.4), which points to **a very large growth effect of telecommunications investment**. It is more than likely that this is an **overestimation**, similar to the results of Aschauer (1989a, 1990) for public capital infrastructure investment.

Bebee and Gilling (1976) used data for 29 developed and undeveloped countries for the year 1970 to regress an index of economic performance on the inverse of a development support factor index and the inverse of an index for telephone use and availability. He hypothesized that per capita GDP of the secondary and tertiary sectors of the economy is a function of certain sector-specific developmental factors which are catalyzed by telephone use and availability. He found a **very strong** positive relationship between the economic performance and telephone indices.

Hardy (1980) used panel data to regress gross domestic product per capita & per capita energy consumption (indicators of economic development) on telephones per million people. A **logarithmic transformation** was used to correct for skewness. In order to test for causality, he lagged the data one year in time. The telecommunications coefficient was significant for GDP per capita and energy consumption regressions. However, he obtained different results for the economic development measures when he considered the influence of **business telephones**. Also, when the regression is estimated separately for developed and developing countries the results are not significant (maybe due to the reduced sample size of range of the variables).

Norton (1992) tested Leff's (1984) hypothesis that telecommunication infrastructure improvements decrease transaction costs. He found that there is indeed a significant and positive effect.

Dhokalia and Harlam (1994) performed their analysis at the US state level. Cross-sectional analysis and separately a lag-model (two time periods, 1985 and 1990: the relationships between availability of resources (1985 data) and economic development (1990 data)). They carried out both a simple regression and a multiple regression analysis. They also incorporated **other resources** in order to carry out a comparative analysis: telecommunication, physical infrastructure such as roads and bridges, human capital through education, and energy. Dependent variables are Average annual pay/per capita income. Independent variables are Telecommunications (Number of business access lines per employee), Education (In dollars; public school expenditure per capita), Physical infrastructure (Municipal and rural highway per square mile of land area), Energy (Energy consumption (in BTU) by the commercial sector per employee). There is a **significant positive** influence found. This relationship was found to be significant in the simple regression as well as in the multiple regression. This in the analysis using 1990 data as well as the lag model using 1985 data. When using average annual pay instead of per capita income, the positive relationship is still significant.

Madden and Savage (1998) analyzed the impact of telecommunications investment in transitional economies in Central and Eastern Europe. This region was chosen because Jipp (1963) found that any positive relationship between investment and economic activity is more significant for lower income economies. They estimated an **aggregate** and **sectoral** economic growth model. The Rate of growth of real GDP per capita was used as an indicator for economic growth and the number of the share of telecommunications investment in GDP (TEL) was used as an indicator for telecom investment. The coefficient for TEL is found to be positive and significant. When the telecom investment share of GDP increases with 1%, the rate of growth of real GDP per capita increases by 13%-15% (this is for the aggregate economic growth model). They also tested for **Granger Causality**. Overall, there appears to be two-way, or mutual precedence between telecommunications investment and real economic growth at the aggregate level

Datta and Agarwal (2004) found lines per 100 inhabitants to be significant as a predictor for rate of growth of real GDP per capita. It is a **dynamic model**: it takes into account the correlation between previous and subsequent values of growth. They allowed for **country fixed effects**: a variable intercept was included in the model. The **square of the telecom variable**, was included to study the nature of returns to scale to telecom investment. This relationship was characterized by **diminishing returns.** Causality was tested by means of inclusion of lagged values of the telecom variable. The causality runs **unilaterally** from telecom to economic growth.

Comparison

Most regressions in this section are conducted using **panel data** (cross-sectional time series data) (**CCITT group, 1968; Hardy, 1980; Datta and Agarwal, 2004**). Panel data is said to have many advantages over cross-sectional or time series data: they allow to account for heterogeneity between subjects, have more variability, more degrees of freedom, less collinearity among variables and more efficiency, are better able to investigate the dynamics of change etc. (**Baltagi, 1995**).

A majority of studies define a **production function** in which telecommunication is either assumed to be a factor of production or a 'technology' factor affecting productivity (comparable with **total factor productivity** in the Cobb Douglas production function). Datta and Agarwal, 2004 modeled their regression function on a standard production function. The advantage of using such a model specification is the sheer amount of theory underlying classical production functions and the inherent simplicity of the concept.

In some cases, the research focus has been on investment in a single developmental input – physical or telecommunications infrastructure – in isolation from other resources required for development (**CCITT group, 1968**; **Hardy, 1980**; **Madden and Savage, 1998**). This allows the implementation of a **simple** (linear) regression model. One objection is the existence of an **omitted-variable bias**, which violates the assumption that the model should be correctly specified. Another is the fact that policy makers have to make trade-offs when carrying out investment decisions so there should be a means of **comparative analysis**. Dhokalia and Harlam, 1994 included next to telecommunications (Number of business access lines per employee), the following independent variables in their regression: Education (In dollars; public school expenditure per capita), Physical infrastructure (Municipal and rural highway per square mile of land area) and Energy (Energy consumption (in BTU) by the commercial sector per employee). Bebee and Gilling, 1976 included an index of development support factor (which for example includes the quality of manpower) in their set of independent variables.

When simply carrying out a multiple regression analysis, problems of **reverse causality** can surface: the beta coefficient of the telecom variable is highly significant but its not clear in which direction (if not both) causality runs. Several (partial) solutions can be implemented to overcome this issue:

- Include the initial year value of the stock of telephone. A significant and positive coefficient of the initial year stock of telephones is used to confirm that the relationship is not the result of reverse causality (**Norton, 1992**).
- Use lagged values of the telecom variables **(Hardy, 1980)**. Short of data mining, it is a non-trivial exercise to determine the appropriate lag.
- Include lagged values of the telecom variable next to the instantaneous values. If the inclusion of these lagged values improve the prediction of the dependent variable, precedence can be deduced, which gives rise to (unilateral) Granger-causality, as discussed in the next subsection. (**Datta and Agarwal, 2004**)

It must be noted that the above operations are imperfect solutions to the problem of reverse causality. A better technique of addressing reverse causality is provided in the next section 'Causality tests'.

Autoregressive/dynamic model: takes into account the relationship between previous and subsequent values of growth. The lagged dependent variable captures short run autoregressive behaviour of the dependent variable. A recurrent issue in estimating autoregressive models using OLS is the existence of serial correlation between the error term and the explanatory variable Y_{t-1}. Because the Durbin-Watson d statistic is not able to discover serial correlation for autoregressive models, Durbin has proposed an alternative large-sample first-order test, called the **h statistics**. For smaller samples, it has been suggested that the **Breusch-Godfrey (BG)** test is more powerful and therefore preferable to the h statistic. After establishing the existence of autocorrelation, the standard errors can be corrected using the **Newey-West HAC** procedure. Datta and Agarwal, 2004 specified a dynamic model.

Another issue is the existence of **fixed effects**. Ignoring individual '**country-effects**' has been shown to lead to biased results (Islam, 1995). Three solutions exist to account for the heterogeneity between subjects in a panel data sample:

- The Fixed Effect Least-Squares Dummy Variable (LSDV) Model: allow each entity to have its own intercept value
- The Fixed-Effect Within-Group Estimator: express values of the independent and dependent values as deviations from their mean
- The Random Effects Model (REM): express the ignorance of the differing intercept values by absorbing the term in the disturbance term

Datta and Agarwal, 2004 implemented the Fixed Effect Least-Squares Dummy Variable (LSDV) model by attributing a variable intercept term to each country in the panel.

The **square of the telecom variable** can be included to study the nature of returns to scale to telecom investment (**Datta and Agarwal, 2004**). This results in a **non-linear** multiple regression model.

Discussion

First and foremost, while multiple regression analysis can model the relationship between multiple variables, it is unable to make any inference of causality between the dependent and independent variables.

An ordinary multiple regression analysis is also **not** advised in situations where the direction of causality is theorized to be in the opposite direction or bilateral. However, as discussed in the previous section, several partial solution exists (as in Norton, 1992; Hardy, 1980 and Datta and Agarwal, 2004) but none is sufficient.

Furthermore, the relationship found could be **spurious** because the time series properties of the data are not taken into account. This means that a relationship may be found when in fact there exist none. Granger and Newbold (1974) showed that standard t- and F-tests are misleading when the underlying timeseries is not stationary. Lee and Gholami (2001) remarked that this potential issue is often neglected or ignored. A solution to avoid the spurious regression problem is cointegration estimation.

2.5.2. Causality tests

Granger Causality was first proposed in 1969 in order to provide a better test for causality than the more traditional regression analysis, which only reflects association (**Granger, 1969)**. A variable X is said to Granger-cause a variable Y if a prediction using the histories of X and Y outperforms a prediction using solely the history of Y. Mathematically, this can be shown if the expected value of Y conditional of historical values of Y and X is different from the expected value of Y given only historical values of Y:

$$\mathbf{E(Y\,|Y_{t-k}, X_{t-k}) \neq E(Y\,|Y_{t-k})}$$

Another definition was provided by Granger in 1969: if the variance of Y predicted based on the entire universe of information (U) is smaller than the variance of Y predicted using U without X, than one can state that X Granger-causes Y. Mathematically:

$$\boldsymbol{\sigma^2(Y\,|U) < \sigma^2(Y\,|\overline{U-X})}$$

Publication Year	Author(s)	☎	📱	Region	Years	Result
1991	Cronin et al.	X		USA	1958-1988	**Bilateral causality**: telecom investment & GDP
1993b	Cronin et al.	X		USA	1958-1990	**Unilateral causality**: telecom investment to productivity
2001	Dutta	X	X	15 developing and 15 industrialized countries	1970-1993	**Bilateral causality**, but more evidence for telecom investment -> GDP
2003	Chakraborty and Nandi	X		12 Asian countries	1975-2000	**Bilateral causality**: mainline access per 100 and GDP. Low degree of privatization: **unilateral** from telecom to GDP
2004	Cieslik and Kaniewsk	X	X	49 Polish regions	1989-1998	**Unilateral causality**: ratio of telephone subscribers to regional income
2004	Shinjo and Zhang	ICT		38 Japanese	1987-2000	**Unilateral:** productivity growth to ICT

				manufacturing industries, 31 USA industries		(Japan); **Bilateral** (USA)
2005	Chu et al.	ICT		New Zealand	1987-2001	**Unilateral**: from ICT to real GDP
2005	Beil et al.	X	X	US telecom firms	1947-1996	**Unilateral**: GDP to telecom investment

Cronin et al. (1991) used two measures of overall US economic activity: the sum of output of 432 industries, and the annual gross national product (adjusted for inflation). Telecommunications investment was obtained as a specially constructed time series using investment in telecom structures as reported by the Bureau of Economic Analysis and the non-consumer goods output of telecom equipment manufacturers. They used the Granger test, Sims test and Modified Sims test but dropped the Sims test because of its misleading estimates in small samples. A **two-year lag** model was used and an apparent trend was removed by differencing. They found a bilateral relationship between both measures of economic activity and telecommunications investment, which led them to conclude that there is an apparent feedback process between both.

In 1993, Cronin et al. (1993) performed a similar analysis using the same telecommunications investment data (expanded to 1990), but analyzing the relationship between telecom investment and productivity. Three measures of multifactor productivity were utilized: private non-farm business productivity, private business productivity and manufacturing productivity. The Granger and Modified Sims test were chosen for similar reasons as in Cronin et al. (1991), but three different lags were examined: two-, four- and six- year lags. This resulted in 2 (Granger & Modified Sims) multiplied by 3 (3 different lags) multiplied by 3 (three measures of productivity), equals 18 tests. Of these 18 tests, the null hypothesis of a unilateral relationship between telecom investment and productivity was rejected at the 90% confidence level in 16 of those tests, and at the 95% (of higher) confidence level in 11 of those tests. As a result, Cronin et al. conclude that there is an obvious unilateral causality running from telecom investment to productivity.

Dutta (2001) examined two series of countries: developing and industrialized. For each of these countries he had 2 sets of observations: per capita GDP in 1987 dollars & telephones per 100 persons; and total GDP in 1987 dollars and total number of telephones. Because of the emergence of wireless telephony, he included cellular subscribers for countries and years for which the data was available. Additionally, he performed the analysis with the raw time series and with a second set after applying the logarithmic transformation. Dutta only used the direct Granger test for his analysis. He found *reasonable* evidence for causality running from telecom investment to economic activity while the evidence for the inverse causal relationship is substantially weaker. He found no difference in pattern between developing and industrialized countries.

Chakraborty and Nandi (2003) examine the relationship between mainline access per 100 inhabitants and annual real GDP measured in 1995 dollars. Using a direct granger test with k=2 lags, they found a bidirectional causality between both for the long and short term. Additionally, they created country subgroups based on the degree of privatization in the telecommunications sector (high and low degree of privatization). They find that for countries with a high degree of privatization, causality is bilateral in the short and long term. On the other hand, for countries with a low degree of privatization, causality runs only from telecom to GDP. This last result shows that an appropriate regulatory context is needed to translate demand for telecom services (spawned by income increases) to the supply of those services.

Cieslik and Kaniewsk (2004) investigated the causality between number of telephone subscribers per 100 000 inhabitants and retail sales per worker in 1997 prices (as proxy for regional income per worker). They then used Granger's (1969) causality test with Akaike's (1970) final prediction error (FPE) criterion. They found that the higher income in richer regions is caused by a higher teledensity, in other words, they find unilateral causality between telecommunications and income.

Shinjo and Zhang (2004) test the causality between productivity growth and ICT for Japanese and American (USA) industries. Lag lengths from 2 to 4 years are used. For Japanese industries, it is found that productivity growth causes ICT investment while for USA industries there is a clear bilateral relationship between both.

Chu et al. (2005) used Granger causality test on three sets of New Zealand ICT variables and real GDP. The ICT variables have relation to specific industries: the two most ICT-intensive industries were found to be the communication service sector and the finance and insurance sector. A third measure adds these two sectors together. After applying unit root tests and applying cointegration estimation, they found unilateral causality from ICT to GDP.

Beil et al. (2005) used 50 years of data on real GDP and real US Telecom firm investment. They found strong evidence that real GDP causes telecommunications investment.

Comparison

Three Granger-causality tests exist for the detection of **causality** :

- Direct Granger test
- Sims test (least used)
- Modified Sims test

Multivariable causality can be investigated using the technique of **vector autoregression** (**VAR**), but this is also out of scope of this thesis.

4 potential outcomes can be found when performing a Granger causality on variables X and Y:

- Unidirectional causality from X to Y
- Unidirectional causality from Y to X
- Bilateral causality between X and Y
- Independence

When trying to find the causal relationship between two variables, the **lag** between source and result is of utmost importance. Standard practice looks at the Akaike Information Criterion (AIC) and chooses a lag length which minimizes this criterion. The procedure is as follows: one picks a maximum lag length (M) and runs a regression for all possible lag combinations (MxM=M^2). One then selects the lag combination which minimizes the information criterion.

Causality tests assume that the underlying time series are **stationary**: covariance stationary or weakly stationary: its mean and variance are constant over time, the covariance between two time periods depends only on the gap between the two time periods and not the time at which the covariance is computed. A stationary time series is characterized by **mean reversion**: it will always revert to its mean and fluctuations will have a constant amplitude. Non-stationary stochastic processes, a well-known example being the **random walk model** (RWM).

There exist several methods (graphical and analytical) to look for non-stationarity:

- Graphical analysis
- Autocorrelation Function (ACF) and Correlogram (Q statistic & Ljung-Box (LB) statistic)
- <u>**Unit root tests**</u>
    - Dickey-Fuller (DF) test
    - Augmented Dickey-Fuller (ADF) test – does not assume that the error term u_t is uncorrelated
    - Phillips-Perron (PP) Unit Root Tests – take care of the serial correlation in the error terms without adding lagged difference terms

A simple possible method of correcting non-stationarity consists of **differencing** the problematic time series.

1. The direct Granger test

The direct Granger test allows to detect the presence and direction of Granger causality. Each variable is regressed on the lagged versions of itself and the other variable. Mathematically:

$$\boldsymbol{Y_t = \sum_{j=1}^{m} \alpha_j Y_{t-j} + \sum_{i=1}^{n} \beta_i X_{t-i} + D_t + \varepsilon_t}$$

If β=0 for all i≥1, the null hypothesis of non-causality cannot be rejected. To determine this, an F-test should be carried out where H_0: $\beta_1= \beta_2=\ldots= \beta_n$.

More **powerful** than the Sims and Modified Sims tests.

Because of overfitting of the model (to remove autodependencies), Granger coefficients are <u>not optimal</u> (they are unbiased but not most efficient). They should therefore not be used for effect sizes but only to test for significance.

It is crucial to include the appropriate n

umber of lagged terms. The **Akaike or Schwartz information criterion** can be used to find the answer.

2. The Sims test

In contrast to the direct granger test, this test is able to detect the **direction** of causality. The Sims test assumes that both times series being tested are jointly <u>covariance-stationary</u>. A series is covariance-stationary if its mean nor its autocovariance (variance of the variable against a lagged version of that variable) is a function of the time t.

$$X_t = au_t + bv_t$$

$$Y_t = cu_t + dv_t$$

Ut and vt are uncorrelated white noise error terms having variance = 1. For t < 0, a,b,c and d are equal to 0.

Subsequently, Y is regressed on past and future values of X. If the coefficients of the future values of X do not significantly differ from 0, granger causality from X->Y can be detected. Because this test depends on the accuracy of a F-test, the assumption of no serial correlation in the residuals has to hold. In order to guarantee this, all variables in the regression will be converted to the natural logarithm and prefiltered using the filter: $1 - 1.5L + 0.5625L^2$

So each variable (X_t) will be transformed as such: $X_t - 1.5X_{t-1} + 0.5625X_{t-2}$.

After this filtering, the following regression is run:

$$\tilde{Y}_t = \sum_{i=-n}^{m} \beta_i \tilde{X}_{t-1} + D_t + \varepsilon_t$$

The null hypothesis of no causality being: $H_0\colon \beta_{-1} = \beta_{-2} = \cdots = \beta_{-n} = 0$

A Wald test is used to compare the restricted and unrestricted versions which results in an F test which enables us to test the null hypothesis.

It has to be noted that the Sims test has **the least power** and has serious flaws of **spurious regression**, which makes it the least used in empirical studies.

3. The modified Sims test

A modified version of the Sims test was proposed by Geweke, Meese and Dent (1983). It is based on the OLS estimation of

$$Y_t = \sum_{j=-m}^{m} \theta_j X_{t-j} + \sum_{i=1}^{n} \delta_i Y_{t-1} + \mu + v_t$$

μ is the deterministic term, v_t is the stochastic error term. In contrast with the 'normal' Sims test, there is dealt with serial correlation by including lagged values of Y_t in the regression.

As before, one can test that Y_t causes X by testing that $\theta_j = 0$ for j=-1,-2,…-m. The restricted and unrestricted versions are tested, which results in an F statistic which is then compared to the critical values.

Discussion

Because of causality being an inherent philosophical concept and the **post hoc ergo propter hoc** fallacy ("after this, therefore because of this", or in other words: precedence doesn't necessary imply causation), one should be wary to equate Granger-causality with true causality. Because of this, other terms are generally preferred: **precedence**, **predictive causality** or **Granger-causality** (**G-causality**).

2.5.3. Simultaneous-Equations model

Simultaneous-Equations models are useful when the unidirectional cause-and-effect relationship (from X's to Y) as found in the standard multiple regression model is not applicable to the problem at hand. Instead, different equations (one for each endogenous variable) are estimated to capture the two-way or simultaneous relationship between Y and the X's. An example system of equations with two endogenous variables is shown mathematically:

$$\boldsymbol{Y_{1i} = \beta_{10} + \beta_{11} Y_{2i} + \gamma_{11} X_{1i} + u_{1i}}$$

$$\boldsymbol{Y_{2i} = \beta_{20} + \beta_{21} Y_{1i} + \gamma_{21} X_{1i} + u_{2i}}$$

Publication Year	Author(s)	☎	📱	Region	Years	Results
1996	Roller and Waverman	X		21 OECD countries	1971-1990	Positive, not significant: **0.26** elasticity; Critical mass of 24% in non-linear model
2004	Sridhar and Sridhar	X	X	63 developing countries	1990-2001	All telephone lines: 0.10; Landlines: 0.14; Mobile phones: 0.007
2008	Shiu and Lam	X	X	22 regions of China	1978-2004	
2010	Gruber and Koutroumpis		X	192 countries	1990-2007	

The ordinary least squares (OLS) method cannot be applied because some of the explanatory variables are correlated with the disturbance term and thus violating one of the assumptions of the Classical Normal Regression Model. This method of estimation would result in **inconsistent** estimates (even if the sample size increases indefinitely, there is no convergence to the true population value).

In **Roller and Waverman (1996)**, the interrelationship between GDP and telecommunication infrastructure is represented by the following set of equations (note that the variables GDP, PEN and TTI are present on both sides of different equations):

$$\log\left(GDP_{it}\right) = a_{0i} + a_1 \log\left(K_{it}\right) + a_2 \log\left(TLF_{it}\right) + a_3 PEN_{it} + a_4 t + \varepsilon_{it}^1$$

$$PEN_{it} + WL_{it} = b_0 + b_1 \log(GDP_{it}) + b_2 \log(TELP_{it}) + \varepsilon_{it}^2$$

$$\begin{aligned}\log(TTI_{it}) = c_0 + c_1 \log\left(GA_{it}\right) + c_2 GD_{it} + c_3 (1 - USCAN) \cdot WL_{it} \\ + c_4 (1 - USCAN) \log(TELP_{it}) + c_5 USCAN \cdot \log(TELP) + \varepsilon_{it}^3\end{aligned}$$

$$PEN_{it} - PEN_{i,t-1} = d_0 + d_1 \log(TTI_{i,t-1}) + d_2 \log\left(GA\right) \varepsilon_{it}^4$$

They use a classical growth/production function (equation 1) together with an endogenous micro-model of supply and demand for the telecommunications sector (equation 2 and 3). Equation 4 translates yearly telecommunication investment into total, accumulated telecommunications infrastructure.

The production function has income (GDP) as a function of capital (K), labour (TLF) and telecom penetration (PEN). For the demand equation, in order to include the **effective demand**, the sum of existing teledensity (PEN) and the waiting list (WL) is taken. This is an important correction over solely using existing teledensitiy because waiting lists can be very substantial, especially in developing regions of the world. Effective demand is a function of income (GDP) and the price of telecom (TELP). Real investment in telecom infrastructure (TTI) is expressed as a function of the geographic area (GA), the real government deficit (GD), the waiting list (WL) and the price of telecom (TELP). The difference in penetration levels between periods is a function of the real investment in telecom infrastructure (TTI) in the previous period and the geographic area (GA).

They found the PEN variable to be positive and significant in the production function equation. A one percent increase in the penetration rate is estimated to increase economic growth by 0.55% on average. This result seems to be an overestimation (in the style of Aschauer (1989a, 1990).

Consequently, they correct for fixed effects in the production function by introducing a **variable intercept** per country (The Fixed Effect Least-Squares Dummy Variable (LSDV) Model, as discussed in the Multiple Regression part). As a result, the coefficient of the penetration rate in the production function is halved, now being 0.26 and is statistically no longer significant!

Next, in order to investigate the existence of nonlinearities and a **critical mass** phenomenon (the impact of telecom investment might only exist when a critical mass of telecommunications stock is accumulated), the square of the PEN variable is added to the production function. The critical mass phenomenon is line with the theory of **network externalities**. The coefficient of PEN squared was found to be positive and significant. The critical mass was calculated to be around a 24% penetration rate.

Sridhar and Sridhar (2004) use a similar set of equations but vary certain variables to come up with different model variations.

$$\text{Log}\ (\text{GDP}_{it}) = a_{0i} + \text{a}_1 \log\ (K_{it}) + a_2 \log\ (LF_{it}) + a_3 \log\ (TPEN_{it}/MTEL_{it}/CELL_{it}) + a_4 t + \varepsilon^1_{it}\ (1)^6$$

$$\text{Log}\ (\text{TPENWL}_{it}/\text{MTELWL}_{it}/\text{CELL}_{it}) = b_0 + b_1 \log\ (GDPCAP_{it}) + b_2 \log(TREVUSR_{it}/MREVUSR_{it}/CREVUSR_{it}) + \varepsilon^2_{it}$$

$$\text{Log}\ (\text{TPENWL}_{it}/\text{MTELWL}_{it}/\text{CELL}_{it}) = b_0 + b_1 \log\ (GDPCAP_{it}) + b_2 \log(TELP_{it}/MLPRC/E_{it}/CELLPRCE_{it}) + \varepsilon^2_{it}$$

$$\text{Log}\ (\text{TTI}_{it}) = c_0 + c_2 WL_{it} + c_3 \log(TREVUSR_{it})/\log(TELP_{it}) + \varepsilon^3_{it}$$

$$\text{Log}\ (CHGTEL_{it}/CHGMTEL_{it}/CHGCELL_{it}) = d_0 + d_1 \log(TTI_{it}) + \varepsilon^4_{it}$$

Firstly, three different models are being estimated using the total telecom penetration as the sum of main line and cellular teledensity **(TPEN)**, number of main lines per 100 **(MTEL)** and number of cellular subscribers per 100 **(CELL).** The **income function** is further almost identical to that of Roller and Waverman (1996). The first equation is tested with and without variable intercept. For the **demand function**, effective demand also includes the waiting list for total and main line telecom (TPENWL and MTELWL respectively). Two different variations on the model are introduced here by using either the total revenue per user in $US (TPENWL/MTELWL/CELL) or the real monthly subscription in $US (TELP/MLPRCE/CELLPRCE) as an explanatory variable. As in Roller and Waverman's (1996) model, income (here as real GDP per capita GDPCAP) is used as another explanatory variable. The **supply function** features the waiting list (WL) and telecom revenue (or price) (TREVUSR/TELP) as explanatory variables. The **growth in telecom penetrations** is modeled as s function of real annual telecommunications investment.

They also estimate the resulting model **with** and **without** fixed effects, as previously done by Roller and Waverman (1996).

For the all telephone lines' model, the elasticity for teledensity is 0.15 without fixed effects and 0.10 with fixed effects. For the landlines' model, the elasticity is 0.14 with and without fixed effects. On the other hand, for mobile phones this elasticity is estimated at 0.007 (with fixed effects), small but still significant. It has to be noted that the data regarding mobile phones is quite limited because of the unreliability and unavailability of such data for developing countries before 1996.

Gruber and Koutroumpis (2010) feature the most recent, influential application of the simultaneous equation model, first featured in **Roller and Waverman (1996)**. It defines an aggregate production equation, demand equation, supply equation and mobile infrastructure production equation (see infra).

Aggregate Production equation:

$$GDP_{it} = a_1K_{it} + a_2L_{it} + a_3Mob\_Pen_{it} + a_4Urb_{it} + \varepsilon_{1it}$$

Demand equation:

$$Mob\_Pen_{it} = b_1GDPC_{it} + b_2MobPr_{it} + b_3Urb_{it} + \varepsilon_{2it}$$

Supply equation:

$$Mob\_Rev_{it} = c_1MobPr_{it} + c_2GDPC_{it} + c_3HHI_{it} + \varepsilon_{3it}$$

Mobile infrastructure production equation:

$$\Delta Mob\_Pen_{it} = d_1Mob\_Rev_{it} + \varepsilon_{4it}$$

Besides capital (Fixed stock of capital million USD), labour (Population with full or part time work aged 15-64 in thousands) and mobile penetration, Gruber and Koutroumpis also add **urbanization** (Percent of population living in urban area). This regressor was included because of its pronounced effect on growth.

Demand for telecommunication services is modeled as a function of price (price of a standard service for the connection to the network), income (GDP per capita) and urbanization. Urbanization is again included as a predictor because the higher share of the telecom sector in more urbanized areas and the higher willingness of the population to adopt innovative technologies.

The supply equation links telecommunication revenue to price (price of a standard service for the connection to the network), income (GDP per capita) and the Hirschman-Herfindahl (HHI) market concentration index.

Finally, the mobile infrastructure production equation links mobile penetration to mobile revenue. The underlying assumption is that higher mobile revenues speed up investment in the sector and the growth of mobile penetration within the population.

Roller and Waverman (1996), Sridhar and Sridhar (2004) and Gruber and Koutroumpis (2010) use a **non-linear three stage least squares** to estimate their model of equations.

Shiu and Lam (2008) introduce a model with two equations which solely relates GDP to Teledensity (TEL) (and vice versa). Their model can be described as a **dynamic panel data model**: it includes lagged values of both dependent variables on the right-hand side (dynamic or autoregressive) and has a time (178-2004) and spatial (22 regions of China) dimension. A potential weakness of their model is the absence of other potentially relevenat regressors, which can constitute a **specification error**.

$$GDP_{it} = \alpha_1 + \sum_{m=1}^{M} a_m TEL_{i,t-m} + \sum_{m=1}^{M} c_m GDP_{i,t-m} + \mu_i + \eta_t + v_{it} \quad (1)$$

$$TEL_{it} = \alpha_2 + \sum_{n=1}^{N} b_n GDP_{i,t-n} + \sum_{n=1}^{N} d_n TEL_{i,t-n} + \omega_i + \tau_t + e_{it} \quad (2)$$

μ_i and ω_i are the province -specific effects (i being the province-index). η_t and τ_t are the time period dummies (t being the time index). The error terms are represented by v_{it} and e_{it}.

In order to to **exclude** the province-specific effects and be certain of variable stationarity (integration order of 1), the two time series are differenced to obtain the following equations:

$$GDP_{it} - GDP_{i,t-1} = \sum_{m=1}^{M} c_m (GDP_{i,t-m} - GDP_{i,t-m-1}) + \sum_{m=1}^{M} a_m (TEL_{i,t-m} - TEL_{i,t-m-1}) + (v_{it} - v_{i,t-1})$$

$$TEL_{it} - TEL_{i,t-1} = \sum_{n=1}^{N} d_n (TEL_{i,t-n} - TEL_{i,t-n-1}) + \sum_{n=1}^{N} b_n (GDP_{i,t-n} - GDP_{i,t-n-1}) + (w_{it} - w_{i,t-1}).$$

Instead of using 2SLS or 3SLS, Shiu and Lam use the **Generalized Method of Moments (GMM)** to estimator featured in Arellano and Bond (1991) . A discussion of this statistical method falls out of the scope of this thesis but it involves using different lags of the independent and dependent variables to be used as **instrumental variables**.

Comparison

Sridhar and Sridhar (2004) improved on Roller and Waverman's (1996) model by examining the impact of fixed telephony, mobile telephony and the two combined, and is as such a more extended analysis. They also explored the incorporation of different variables compared to Roller and Waverman (1996). On the other hand, they do not include non-linear, critical mass effects.

Discussion

Once **simultaneity** and **fixed effects** are accounted for, little effect is found (cfr. other public infrastructure investment econometric studies). There are clear indications that non-linearities exist, which corresponds with the theory of network externalities.

3. Research Questions

As discussed previously, this thesis attempts to estimate the impact of infrastructural developments on people's livelihoods. This thesis examines two different infrastructural investment areas and two different interpretations of livelihood improvements.

Firstly, two distinct infrastructure types are being examined: telecommunications and electricity generation/consumption. Both are abundantly discussed in the literature but there is still no consensus on the exact impact of its effect on GDP and other developmental indices.

Secondly, two different livelihood factors are analyzed: GDP per capita and Female Genital Cutting (FGC). GDP per capita is a pure income measure while FGC is the result of complex social and historical prerogatives.

Lastly the effects between infrastructure and livelihood are being researched from two perspectives: a macro (data on country-level) and micro (data on the level of the individual) view.

The whole thesis and subsequent analysis is centered around Africa.

3.1. Macro analysis

In order to investigate the causality between infrastructure investment and GDP per capita, a country level perspective is used. South Africa is taken as an example because it is the only country in Africa which has data going back several decennia. Also, when investigating the rise of the mobile handheld in South-Africa, it is very useful that the country possesses quarterly GDP data: using quarterly data, we can gather a lot of datapoints for GDP and mobile penetration although the mobile revolution has occurred quite recently.

We will use the **Modified Sims test** to find the direction of causality between telecommunications and electricity infrastructure and GDP per capita. The statistical program beins used is **EViews**.

3.2. Micro analysis

A micro perspective is used to determine if mobile phone adoption and electricity consumption has the ability to directly promote peoples' wellbeing in social/health terms. In a world's first, DHS and Unicef data is being used to investigate a whole series of different potential variables and their impact on the rate of FGC.

Another technical approach is used to analyze this huge quantity of data: **machine learning**. The program being utilized is **Python with the Pandas and Numpy packages**.

4. Methods and Processes

4.1. Data Sourcing

Mobile Telephony: Unique subscribers (%)

In choosing a metric to represent the massive evolution in telecommunications usage in Africa, one has to be aware of the advantages and disadvantages of each metric.

First and foremost, one should clearly distinguish between **mobile connections** (SIM cards) and **unique mobile** subscribers. GSMA Intelligence, the market research division of the global GSM association, estimated that the number of mobile connections reached 7 billion on April 8th, 2014. This, however, didn't mean that every person on the globe was in possession of a SIM card. Worldwide, the average mobile subscriber has 2 SIM cards, which makes the percentage of people connected to a mobile network roughly 50%. To take an example, Chinese subscribers were found by GSMA to have on average 1.79 SIM card as of Q4 2013. This means that while the total number of connections amounted to 1.25 billion, less than half the population (630 million) had subscribed to a mobile service.

Furthermore, counting mobile phones as a proxy for telecommunications usage is an inherently flawed approach. In developing countries, the phenomenon known as "device sharing" makes it possible to be a mobile subscriber without owning a device. Besides this fact, counting phones would also include all other connected devices (tablets, routers, dongles etc.).

Lastly, one should keep in mind that the concepts "mobile subscribers" and "mobile users" are not equivalent. Mobile users are individuals in range of a mobile network which theoretically can use a mobile device to connect to that network, irrelevant of the fact that they have subscribed to this mobile service or not.

Because of all these points, a choice was made to use the GSMA's "Market penetration (unique subscribers)" metric in order to measure the extent to which telecommunications has been adopted by the population of a country.

Female Genital Cutting (FGC)

In 2010, for selected countries, UNICEF (through its programmes MICS4 and MICS5) and USAID (through DHS) started using synchronized FGC modules while both also began gathering circumcision data for daughters under 15 years of age. Because of the fact that the act of circumcision lies many decades in the past for most interviewed women, the daughter FGC module is of most interest to this study. It allows to detect FGC trends up to the present day and makes it easier to integrate the data from both sources.

The table below shows the most recent datasets (available to the public) incorporating FGC daughter data. The surveys are ranked by year. As can be seen, MICS is generally slower in releasing data to the public.

Programme	Survey Year	Country
DHS	2014	Senegal
DHS	2013-2014	Togo
DHS	2013	Nigeria
DHS	2012-2013	Mali
DHS	2012	Benin
DHS	2012	Guinea
DHS	2011-2012	Ivory Coast
MICS	2011	Ghana
MICS	2011	Mauritania
DHS	2010	Burkina Faso
MICS	2010	Central African Republic
MICS	2010	Sierra Leone

Codebook

The table below comprises the variables which are selected from the datasets listed in table **Y**. It contains the original variable names in the DHS and MICS datasets (***'DHS'*** and ***'MICS')***, the description of the variable (***'Description')***, the variable type of the variable (***'Type'***) and the new name used in this study (***'New Name')***.

DHS	MICS	Description	Type	New Name
V005	WMWEIGHT **(wm)**	Sample weight	Ratio	SAMPLE_WEIGHT
V007	WM6Y **(wm)**	Year of interview	Interval	YEAR_ITV
V010	WB1Y **(wm)**	Respondent's year of birth	Interval	YEAR_BIRTH
V024	HH7 **(wm)**	Region	Nominal	REGION
V130	religion **(wm)**	Religion	Nominal	RELIGION
V155	WB7	Literacy	Nominal	LITERACY
V131	ethnicity **(wm)**	Ethnicity	Nominal	ETHNICITY
V120	HC8B **(hh)**	Has radio	Nominal (Y/N)	HAS_RADIO
V121	HC8C **(hh)**	Has television	Nominal (Y/N)	HAS_TELEVISION
	HC8C **(hh)**	Has a black and white TV	Nominal (Y/N)	HAS_TELEVISION_B&W
	HC8C1 **(hh)**	Has a colour TV	Nominal (Y/N)	HAS_TELEVISION_COLOUR
V153	HC8D **(hh)**	Has telephone (land-line)	Nominal (Y/N)	HAS_TELEPHONE
	HC9B/A **(hh)**	Has mobile Telephone	Nominal (Y/N)	HAS_MOBILE
G100	FG1 **(wm)**	Heard of female circumcision	Nominal (Y/N)	HEARD_FCIRC
G101	FG2 **(wm)**	Heard of female circumcision (probed)	Nominal (Y/N)	HEARD_FCIRC_PROB
G102	FG3 **(wm)**	Circumcised	Nominal (Y/N)	FCIRC
G103	FG4 **(wm)**	Flesh removed	Nominal (Y/N)	FCIRC_FLESH
G104	FG5 **(wm)**	Just nicked	Nominal (Y/N)	FCIRC_NICKED
G105	FG6 **(wm)**	Sewn closed	Nominal (Y/N)	FCIRC_SEWN
G106	FG7 **(wm)**	Respondent's age at circumcision	Ratio	AGE_FCIRC
G107	FG8 **(wm)**	Who performed circumcision	Nominal	FCIRC_PRACT
G119	FG22 **(wm)**	FCG: continue or be stopped	Nominal (Y/N)	FCIRC_CONTINUE

B2$IDX		Child's year of birth	Interval	CHILD_YEAR_BIRTH_IDX
B4$IDX		Child’s sex	Nominal	CHILD_SEX_IDX
B8$IDX		Child’s age	Ratio	CHILD_AGE_IDX
B5$IDX		Child is alive	Nominal (Y/N)	CHILD_ALIVE_IDX
B6$IDX		Child's age at death	Ratio	CHILD_AGE_DEATH_IDX
(computed)	FG13 **(fg)**	Daughter’s age	Ratio	GIRL_AGE
GIDX$IDX		Daughter's index to birth history	Ratio	GIRL
G121$IDX	FG3 **(fg)**	Is daughter circumcised	Nominal (Y/N)	GIRL_FCIRC
G122$IDX	FG16 **(fg)**	Daughter's age at circumcision	Ratio	GIRL_AGE_FCIRC
	FG16U **(fg)**	Unit daughter’s age at FGC	Nominal	GIRL_AGE_FCIRC_UNIT
	FG17 **(fg)**	Daughter’s genital area flesh removed	Nominal (Y/N)	GIRL_FCIRC_FLESH
	FG18 **(fg)**	Daughter’s genital area nicked	Nominal (Y/N)	GIRL_FCIRC_NICKED
G123$IDX	FG19 **(fg)**	Daughter's genital area sewn closed?	Nominal (Y/N)	GIRL_FCIRC_SEWN
G124$IDX	FG20 **(fg)**	Daughter - who performed circumcision	Nominal	GIRL_FCIRC_PRACT

Methods and Processes

Because of the intricate difficulties involved in processing the raw data, the data analytics module **Pandas** for Python is used. The data was sourced as **SPSS (.sav)** files but these are transformed into **Comma Separated Values (CSV)** files in order to load the data more swiftly into memory. The codebook shown in the previous section (*3.3.1. Codebook)* lists all variables which will be used in subsequent analyses. A modified version of the codebook is used as a python *dictionary* to select and rename the desired variables in all datasets. The subsequent processing steps are quite different for DHS versus MICS databases and as such will be discussed separately below:

DHS

In the DHS surveys each female respondent was asked general information related to her children (all children ever born to the woman, dead or alive at the moment of questioning). Each question **for** each child is coded as a separate variable (eg. Age of Child 3, Sex of Child 4 etc.). All women were also posed questions about the FGC status of their daughters. For these questions, the answers for each daughter were also coded as separate variables. Linking these questions to the initial questions asked for all children, is a variable (for each daughter) ‘**Index to birth history**’ **(GIDX$Index)**. Ideally, we would like to obtain one entry for each girl including age and other general information.

In order to get to this result, the age for each girl is obtained using the ‘Index to birth history’ ‘**key**’: from **B8$Index** (*Child’s age)*, **GIRL_AGE** (*Daughter’s age)* is computed. Next, the FGC questions per daughter are **stacked** (which means they are added as separate entries instead of variables). The information related to the child’s mother is retained for each daughter.

MICS

UNICEF delivers a separate dataset for each unit of study (wife, husband, children, family etc.) or potential research area (FGC of daughters, Use of mosquito nets etc.). For the purpose of this paper, the datasets **women** (wm)**, household** (hh) and **FGC of daughters** (fg) are used. These three databases are merged using a triple component key: **Cluster Number, Household Number** and **Line Number**. 'Line number' is used to distinguish between different women living in the same household. To match the household characteristics to a woman or her (circumcised) daughters, only the cluster number and household number are needed.

Furthermore, while variable names are very similar between countries, there are still small differences, complicating the selection and renaming procedure.

Weighting

DHS and MICS include **sampling weights** in their datasets. Weights are needed when differences in selection (and/or interview) probability occur. This can happen by accident (by refusal of large numbers of eligible individuals) or by design (when quotas are established for certain subgroups to obtain robust statistics about them). Weights are calculated for households and individuals. Calculation of the **raw** weights is carried out in the following way:

$$\text{Household weight} = \frac{1}{\text{HH selection probability}} * \frac{1}{\text{HH response rate (in response group)}}$$

$$\text{Individual weight} = \text{Household weight} * \frac{1}{\text{Individual response rate (in response group)}}$$

These raw weights are then **normalized** (or **standardized**) using the following formula:

$$\text{Normalized weight} = \frac{\text{Sum of number of cases}}{\text{Sum of the raw weights}} * \text{Initial weight} = \frac{\text{Initial weight}}{\text{Average of the raw weights}}$$

One of the characteristics of normalized weights is that they add up to the total number of cases. In the DHS datasets, the normalized weights are represented without decimal point and should be divided by 1,000,000 before use.

Remarks

Sample weights cannot be used when estimating relationships, e.g. regression and correlation coefficients.

The estimation of confidence intervals after incorporation of sample weights is biased. This is because the number of weighted cases is taken instead of the number of cases.

Transformations

$$\text{YEAR\_FCIRC} = \text{YEAR\_ITV} - \text{GIRL\_AGE} + \text{GIRL\_AGE\_FCIRC}$$

- *YEAR_FCIRC: Circumcision year*
- *YEAR_ITV: Year of interview*
- *GIRL_AGE: Girl's age*
- *GIRL_AGE_FCIRC: Girl's age at circumcision*

Unfortunately, in order to calculated the circumcision rate per year, we have to make a reliable estimate of the year for which non-FGC women would have been circumcised if they had undergone the procedure. To produce such an estimate, one should attempt to group all women in similar groups relevant to the act of circumcision. To do so, one can look at factors linked to the act of circumcision, which includes geographical, ethnical, religious and time evolution factors.

Ergo, to obtain such an estimate, the **average age at FGC** was calculated for circumcised females attributed to the same

1) Country,
2) Region,
3) Year of Birth (because average FGC ages evolve over time),
4) Ethnicity *and*
5) Religion

as the not-circumcised female individual.

If combining all variables resulted in a filtered set too small or nonexistent to compute a valid average (for example the girl's family belonged to a particularly obscure religion, which combined with the year of birth, ethnicity and region made comparison to circumcised girls impossible), only the first 4 variables (1-4 in the list published above) were used to compute an average. If this again was not possible, only the 3 first variables were used. This iterative process was continued until a valid average could be computed, which was then used to assign the **hypothetical** girl's age at circumcision.

Limitations of the Data

- MICS only reports religion and ethnicity of the male head of the household. It is assumed that marriages predominantly occur within religious and ethnic groups. Furthermore, religion and ethnicity is not reported for **Mauritania** and **Sierra Leone**.

5. Descriptive Statistics

5.1. Telecommunications

Figure 5 provides a geographic display of the adoption of mobile technology within the African continent. As described in section 4.1., data of GSMA Intelligence on market penetration of unique subscribers is used. The period is the second quarter of 2015.

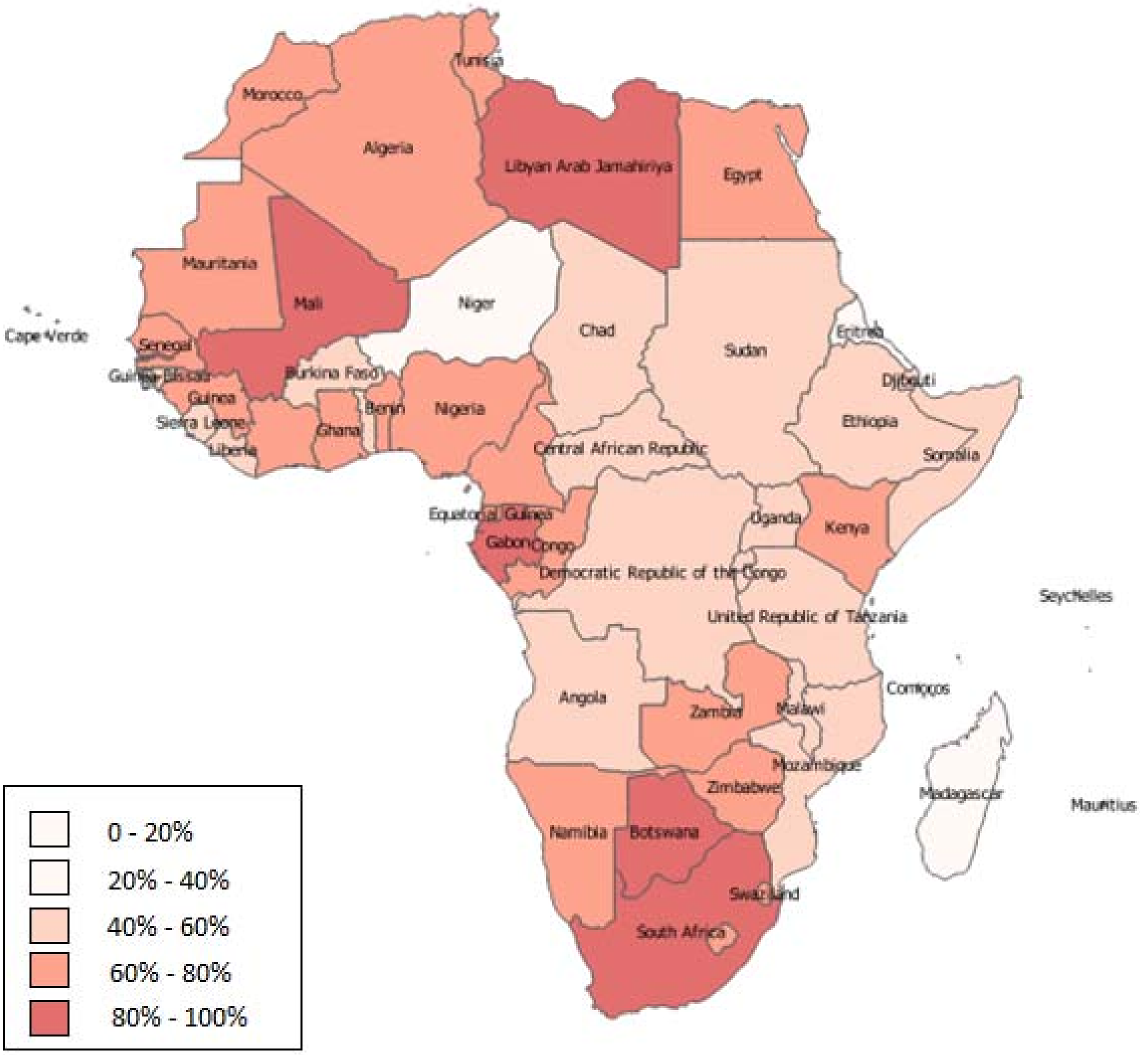


Figure 5: Mobile telephony (market penetration) in Q2 of 2015

Next, the same data is provided for South Africa for the quarters between Q1 2000 and Q3 2015. Astoundingly, during that period, the mobile market penetration has increased from 10% to 70% (which is an increase of 600%). On the other hand, GDP per capita has only increased by 25%.

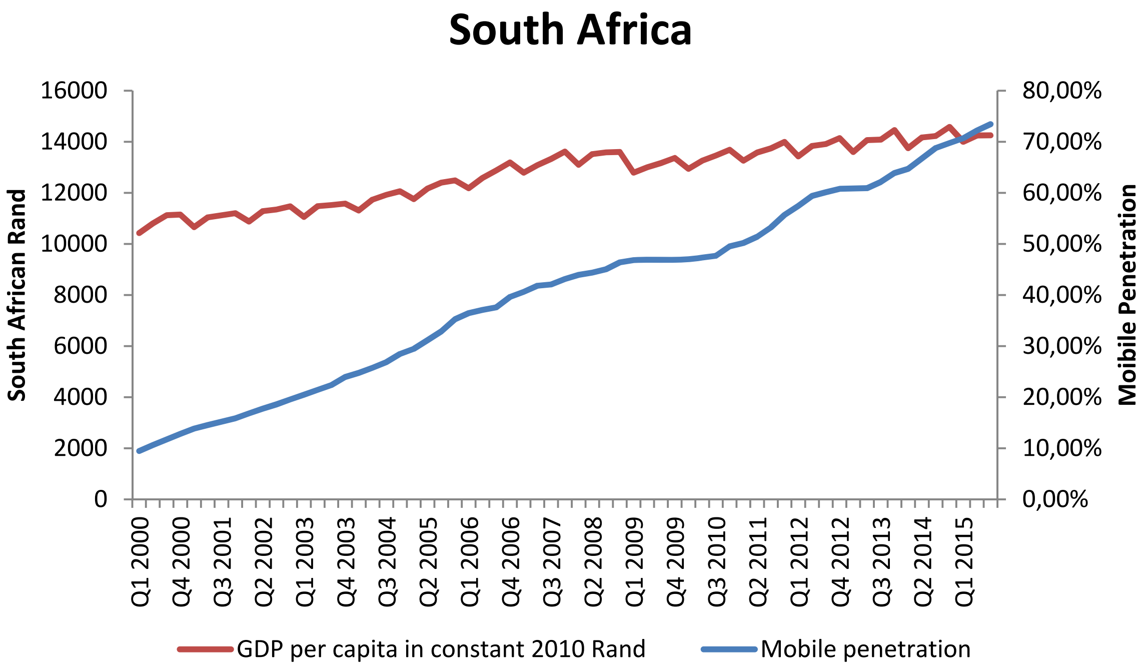


Figure 6: Mobile Market penetration and GDP per Capita for South Africa (Q2 2000 - Q3 2015)

5.2. Electricity Consumption

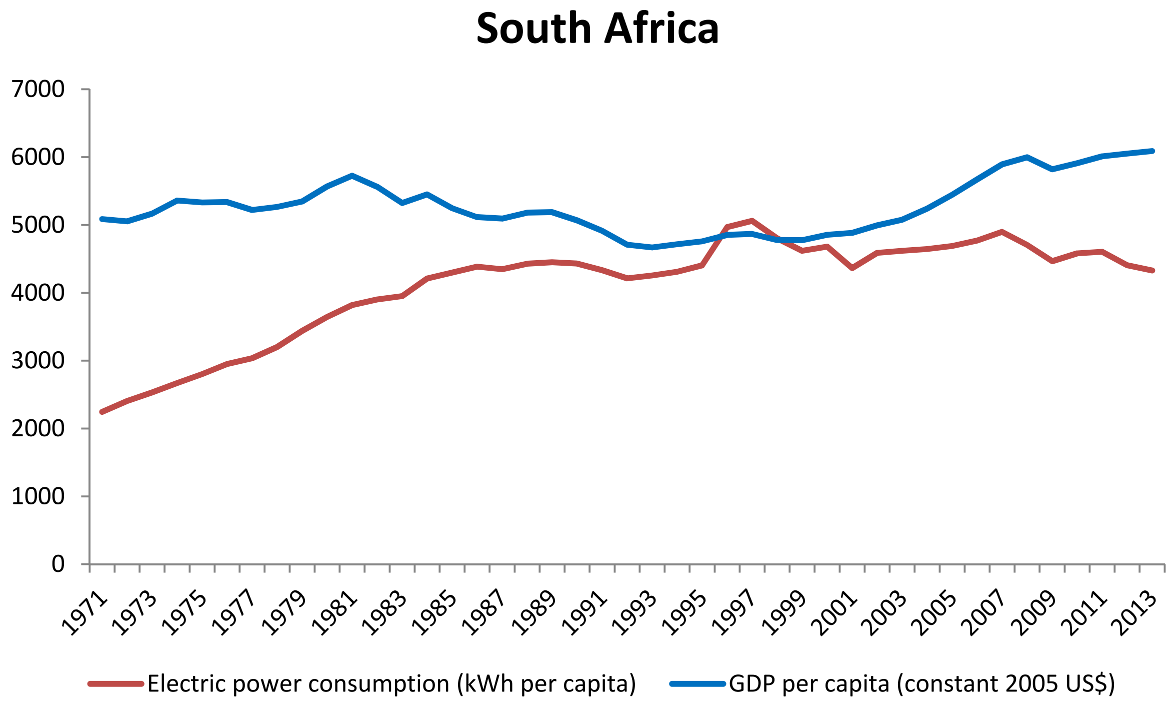


Figure 7: Electricity Consumption and GDP per Capita for South Africa (1971-2014)

5.3. Female Genital Cutting (FGC)

5.3.1. Respondents per country

	DHS		MICS	
Country	**Mothers**	**Daughters**	**Mothers**	**Daughters**
Benin	16 599	11 637		
Burkina Faso	17 087	17 884		
Central African Republic			12 507	17 585
Ghana			10 963	9 079
Guinea	9 142	8 792		
Ivory Coast	10 060	8 904		
Mali	10 424	12 063		
Mauritania			13 657	14 667
Nigeria	38 948	25 223		
Senegal	8 488	8 343		
Sierra Leone			14 066	14 676
Togo	9 480	6 496		

5.3.2. Respondents per country

The next table shows the weighted, cumulative percentage of women who have undergone FGC before and at a particular age**.** The population are all women who (over the course of their life) will eventually undergo FGC. Only women born between 1960 and 1990 (1990 included) are used. This is because >99% of FGC acts happen before the woman reaches 20 years of age (as can be observed from the table) and data may be skewed for women born closer to the moment of surveying.

Age	Benin	Burkina Faso	Central African Republic	Ghana	Guinea	Ivory Coast	Mali	Maurit.	Nigeria	Senegal	Sierra Leone	Togo
0	26%	48%	0%	34%	22%	46%	67%	54%	80%	62%	0%	23%
1	27%	50%	0%	37%	22%	47%	67%	97%	80%	63%	0%	23%
2	27%	53%	0%	39%	22%	48%	71%	98%	81%	65%	1%	24%
3	27%	57%	1%	42%	23%	50%	73%	99%	82%	68%	2%	26%
4	28%	60%	1%	45%	25%	52%	75%	99%	82%	70%	3%	27%
5	31%	68%	3%	49%	30%	58%	80%	99%	83%	75%	6%	31%
6	35%	74%	6%	52%	36%	61%	83%	99%	84%	80%	10%	33%
7	43%	82%	9%	56%	46%	65%	87%	100%	85%	84%	13%	40%
8	55%	86%	17%	62%	56%	70%	90%	100%	87%	87%	20%	44%
9	60%	89%	23%	64%	64%	72%	91%	100%	87%	89%	25%	47%
10	80%	94%	49%	71%	79%	81%	96%	100%	89%	96%	40%	63%
11	82%	95%	54%	76%	83%	83%	97%	100%	89%	97%	44%	66%
12	91%	97%	73%	81%	92%	87%	99%	100%	90%	98%	56%	77%

13	93%	97%	82%	86%	96%	89%	99%	100%	91%	99%	63%	82%
14	94%	98%	88%	88%	97%	91%	99%	100%	92%	99%	72%	84%
15	98%	99%	97%	93%	99%	95%	100%	100%	94%	100%	84%	95%
16	99%	100%	98%	96%	99%	96%	100%	100%	95%	100%	90%	98%
17	99%	100%	99%	96%	99%	98%	100%	100%	95%	100%	92%	98%
18	99%	100%	100%	98%	100%	98%	100%	100%	97%	100%	96%	99%
19	99%	100%	100%	98%	100%	98%	100%	100%	97%	100%	97%	99%
20	100%	100%	100%	99%	100%	99%	100%	100%	98%	100%	99%	99%

Vast differences can be observed over all 12 countries. By the age of 5, only 3% of (to-be) circumcised women have undergone the procedure in the CAR, while this reaches 99% in Mauritania. By the age of 20 the percentage reaches 97%-100% for all countries.

The table is represented graphically infra.

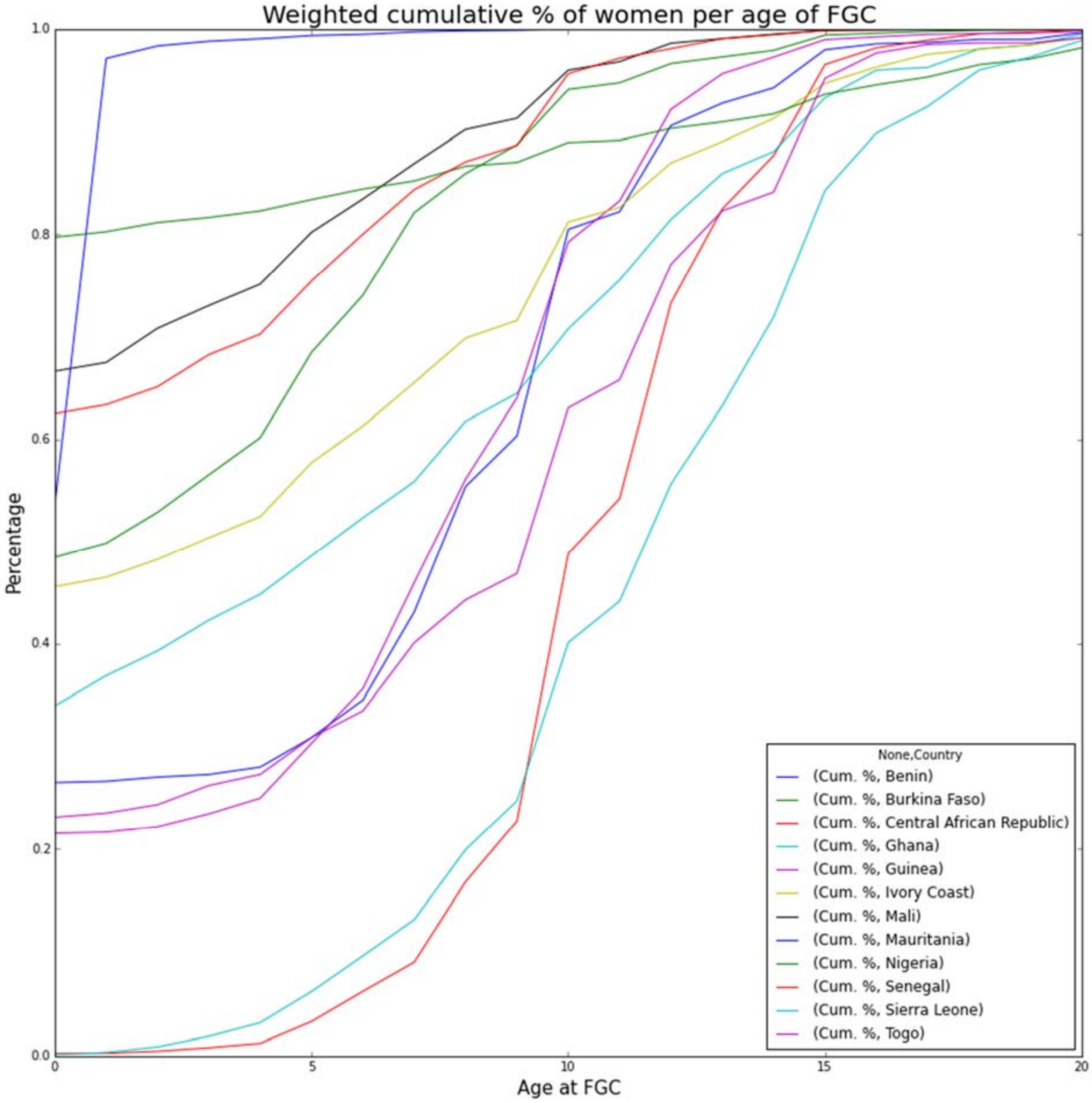


Figure 8: Weighted cumulative percentage of women per age of FGC

Again, using data of women born between 1960 and 1990, the average age at FGC is computed for each country.

Country	N	Average age at FGC (born ’60-’90)
Benin		7.09
Burkina Faso		3.54
Central African Republic		11.00
Ghana		6.40
Guinea		7.11
Ivory Coast		5.18
Mali		2.26

Mauritania	0.16
Nigeria	2.50
Senegal	2.66
Sierra Leone	11.89
Togo	8.22

5.3.3. Historic Evolution FGC incidence per country

Grouped by Year of Birth, Country and Region.

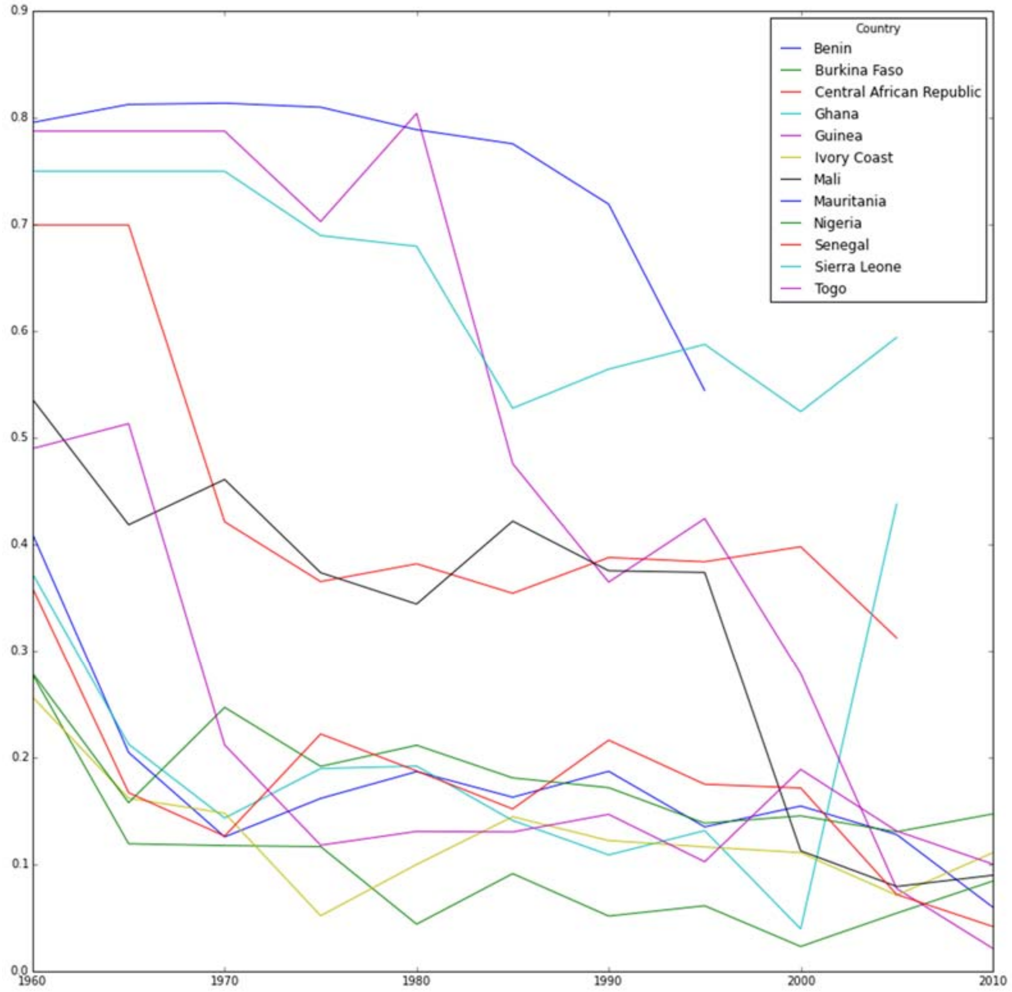


Figure 9: FGC incidence per country

5.3.4. Wealth Index comparison

The wealth index is attempts to provide a quantitative and comparable metric of a household's wealth and is calculated using the household's possession of certain assets, eg. quality of house materials, television, bicycles, access to water and sanitation facilities etc.

It is generated by the DHS and MICS using a statistical procedure known as **principal components analysis**.

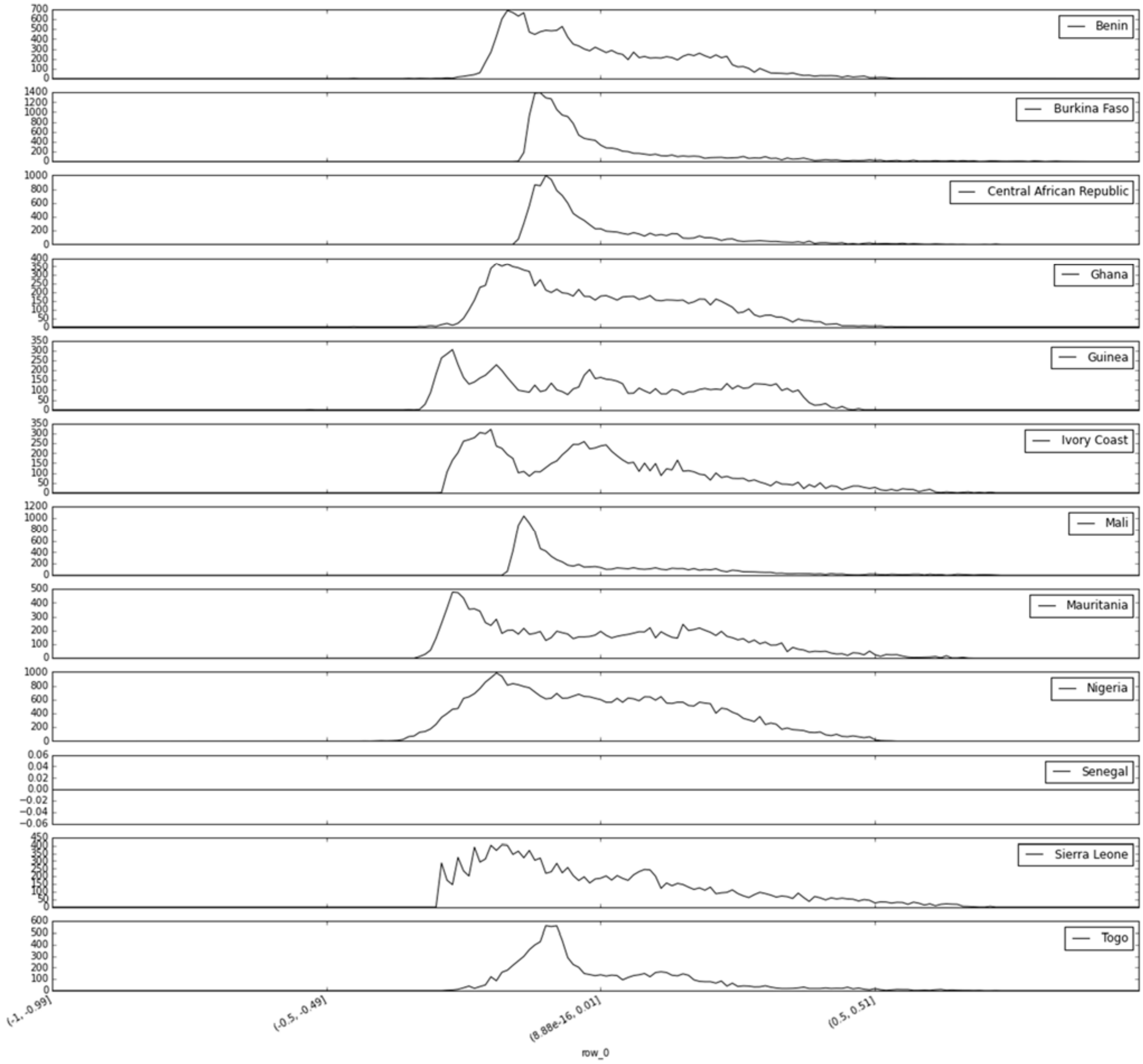


Figure 10: Wealth index distribution per country

5.3.4. Mobile telephony and Mass Media

The following stacked barchart analyzes the ownership of a mobile telephone with regard to the ownership of a radio and/or television.

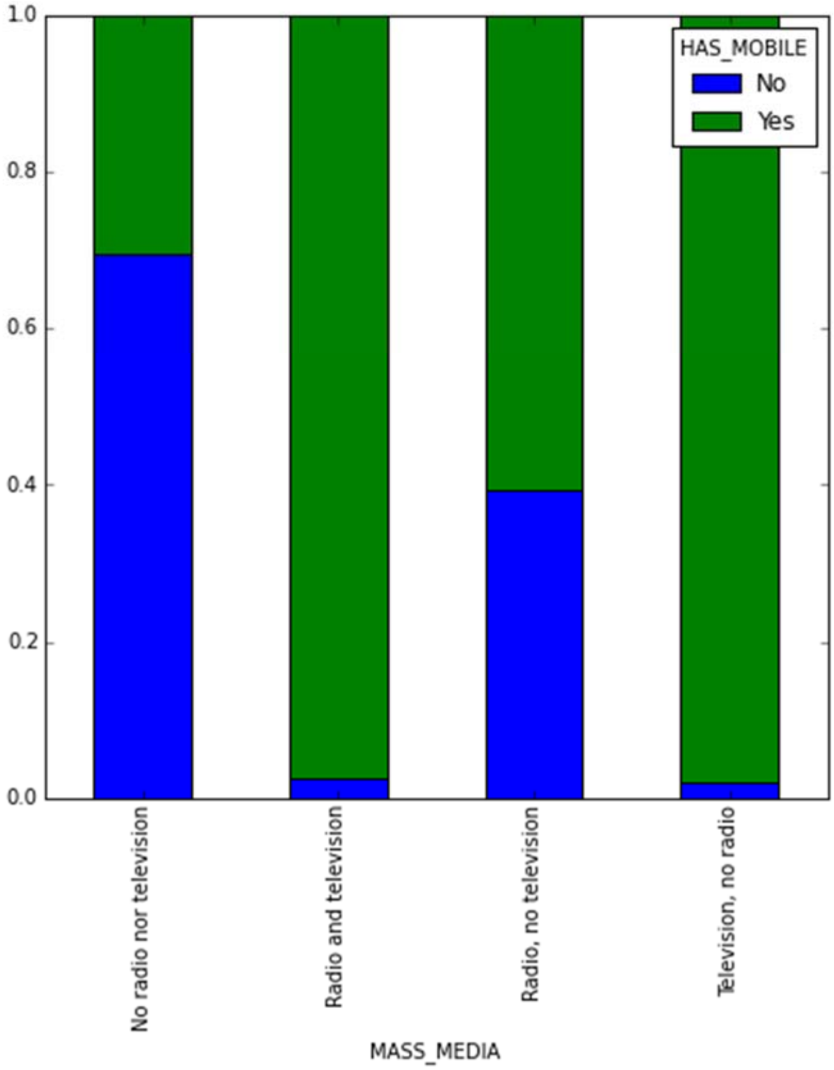


Figure 11: Correlations between mobile telephony and mass media

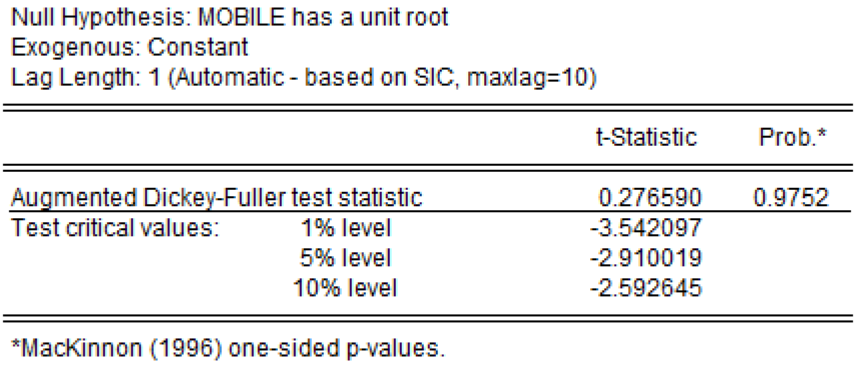
6. Analysis

6.1. Telecommunications

6.1.1. Test for Stationarity

A crucial assumption of the Modified Sims test is that both variables are stationary, i.e. they do not possess a unit root. In order to test this, we use the **Augmented Dickey-Fuller (ADF) test**. This test augments upon the standard Dickey-Fuller test by not assuming that the error term u_t is uncorrelated.

Mobile Market Penetration



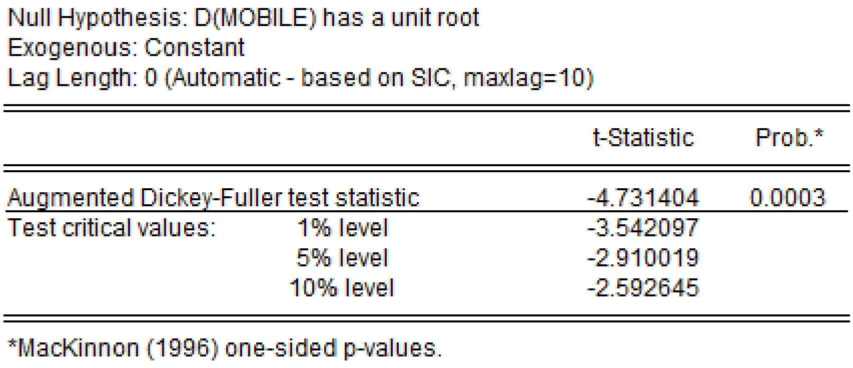
Null Hypothesis: MOBILE has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic - based on SIC, maxlag=10)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		0.276590	0.9752
Test critical values:	1% level	-3.542097	
	5% level	-2.910019	
	10% level	-2.592645	

*MacKinnon (1996) one-sided p-values.

The ADF test statistic has a probability of 0.9752, which means that we **cannot** reject the null hypothesis of non-stationarity (having a unit root) at the 5% significance level.

After **first-order** differencing:



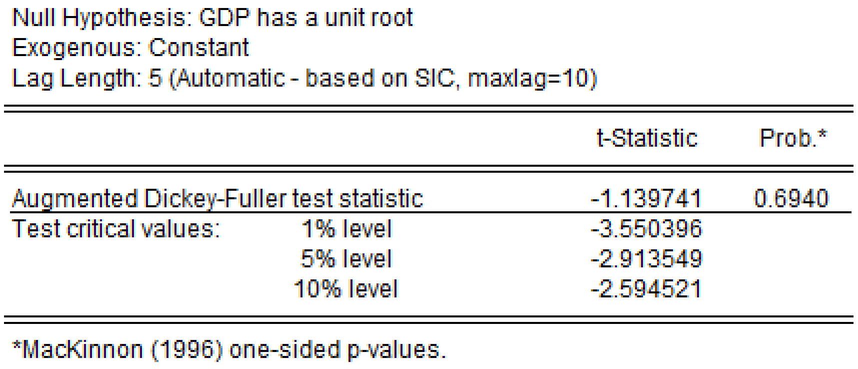
Null Hypothesis: D(MOBILE) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=10)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.731404	0.0003
Test critical values:	1% level	-3.542097	
	5% level	-2.910019	
	10% level	-2.592645	

*MacKinnon (1996) one-sided p-values.

Now, we can reject the null hypothesis at the 5% (and 1%) significance level.

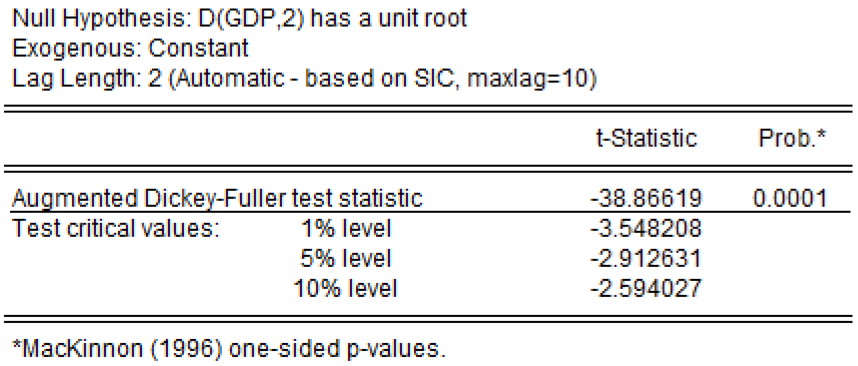
GDP per capita (quarterly, in constant 2010 Rand)



```
Null Hypothesis: GDP has a unit root
Exogenous: Constant
Lag Length: 5 (Automatic - based on SIC, maxlag=10)

                                            t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic      -1.139741     0.6940
Test critical values:       1% level        -3.550396
                            5% level        -2.913549
                            10% level       -2.594521

*MacKinnon (1996) one-sided p-values.
```

With a t-statistic of -1.139741 or a probability value of 0.6940 we cannot reject the null hypothesis of non-stationarity.

After **second-order** differencing:



```
Null Hypothesis: D(GDP,2) has a unit root
Exogenous: Constant
Lag Length: 2 (Automatic - based on SIC, maxlag=10)

                                            t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic      -38.86619     0.0001
Test critical values:       1% level        -3.548208
                            5% level        -2.912631
                            10% level       -2.594027

*MacKinnon (1996) one-sided p-values.
```

Now we can reject the null hypothesis of non-stationarity firmly: 0.0001 < 0.05.

6.1.2. Modified Sims Test

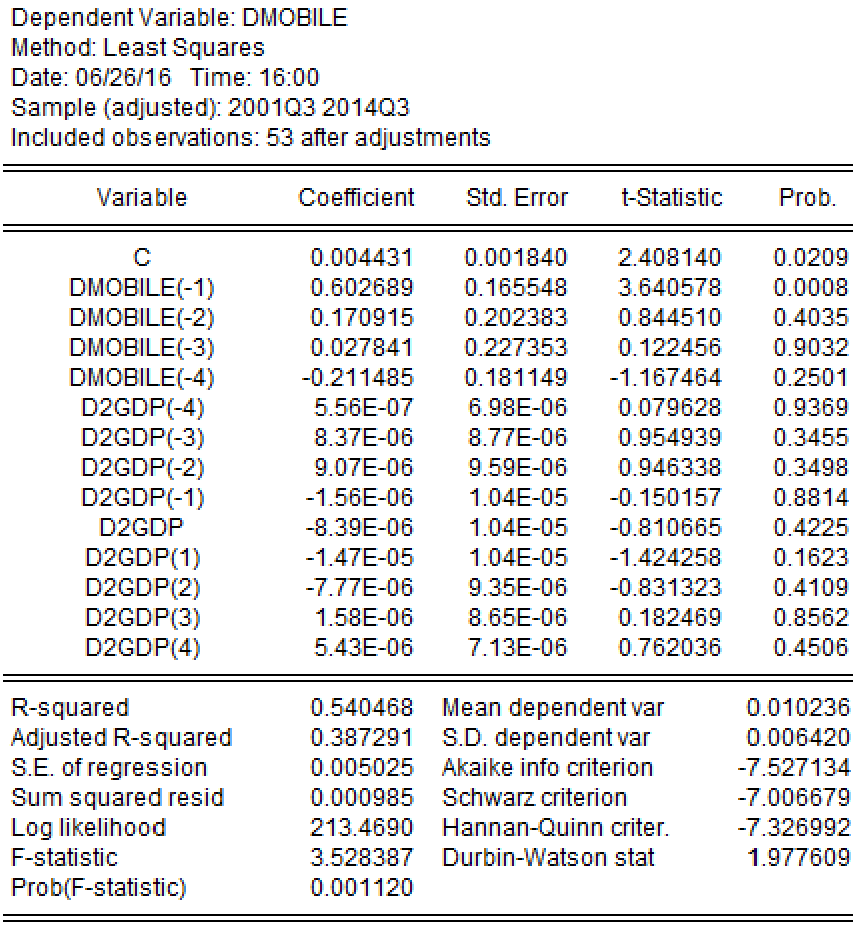
In the following section, we will use the following abbreviations:

- *1st differenced* Mobile Market penetration: **DMobile**
- *2nd differenced* quarterly GDP Per Capita in constant 2010 Rand: **D2GDP**

We will investigate the granger causality from DMobile to D2GDP for a lag period of 4 years and less.

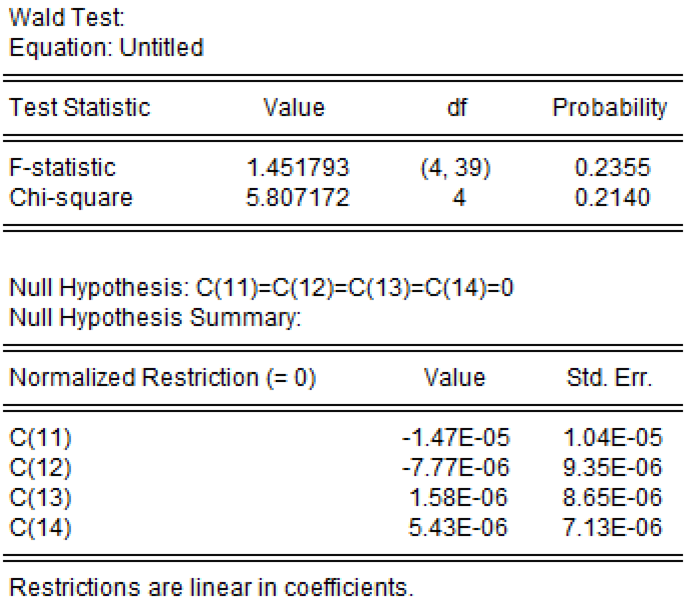
$$DMobile_t = \sum_{j=-4}^{4} \theta_j D2GDP_{t-j} + \sum_{i=1}^{4} \delta_i DMobile_{t-1} + \mu + v_t$$

The **null hypothesis** of this test is of no (Granger) causality from DMobile to D2GDP. After running this regression using OLS, we get the following results:



```
Dependent Variable: DMOBILE
Method: Least Squares
Date: 06/26/16   Time: 16:00
Sample (adjusted): 2001Q3 2014Q3
Included observations: 53 after adjustments
========================================================================
      Variable         Coefficient    Std. Error    t-Statistic    Prob.
========================================================================
         C              0.004431      0.001840      2.408140     0.0209
    DMOBILE(-1)         0.602689      0.165548      3.640578     0.0008
    DMOBILE(-2)         0.170915      0.202383      0.844510     0.4035
    DMOBILE(-3)         0.027841      0.227353      0.122456     0.9032
    DMOBILE(-4)        -0.211485      0.181149     -1.167464     0.2501
     D2GDP(-4)          5.56E-07      6.98E-06      0.079628     0.9369
     D2GDP(-3)          8.37E-06      8.77E-06      0.954939     0.3455
     D2GDP(-2)          9.07E-06      9.59E-06      0.946338     0.3498
     D2GDP(-1)         -1.56E-06      1.04E-05     -0.150157     0.8814
       D2GDP           -8.39E-06      1.04E-05     -0.810665     0.4225
     D2GDP(1)          -1.47E-05      1.04E-05     -1.424258     0.1623
     D2GDP(2)          -7.77E-06      9.35E-06     -0.831323     0.4109
     D2GDP(3)           1.58E-06      8.65E-06      0.182469     0.8562
     D2GDP(4)           5.43E-06      7.13E-06      0.762036     0.4506
========================================================================
R-squared               0.540468    Mean dependent var        0.010236
Adjusted R-squared      0.387291    S.D. dependent var        0.006420
S.E. of regression      0.005025    Akaike info criterion    -7.527134
Sum squared resid       0.000985    Schwarz criterion        -7.006679
Log likelihood          213.4690    Hannan-Quinn criter.     -7.326992
F-statistic             3.528387    Durbin-Watson stat        1.977609
Prob(F-statistic)       0.001120
========================================================================
```

After test, a Wald test for coefficient restriction is run in order to obtain an F statistic and test the null hypothesis that all coefficients on the lead values of DGDP are all equal to zero.

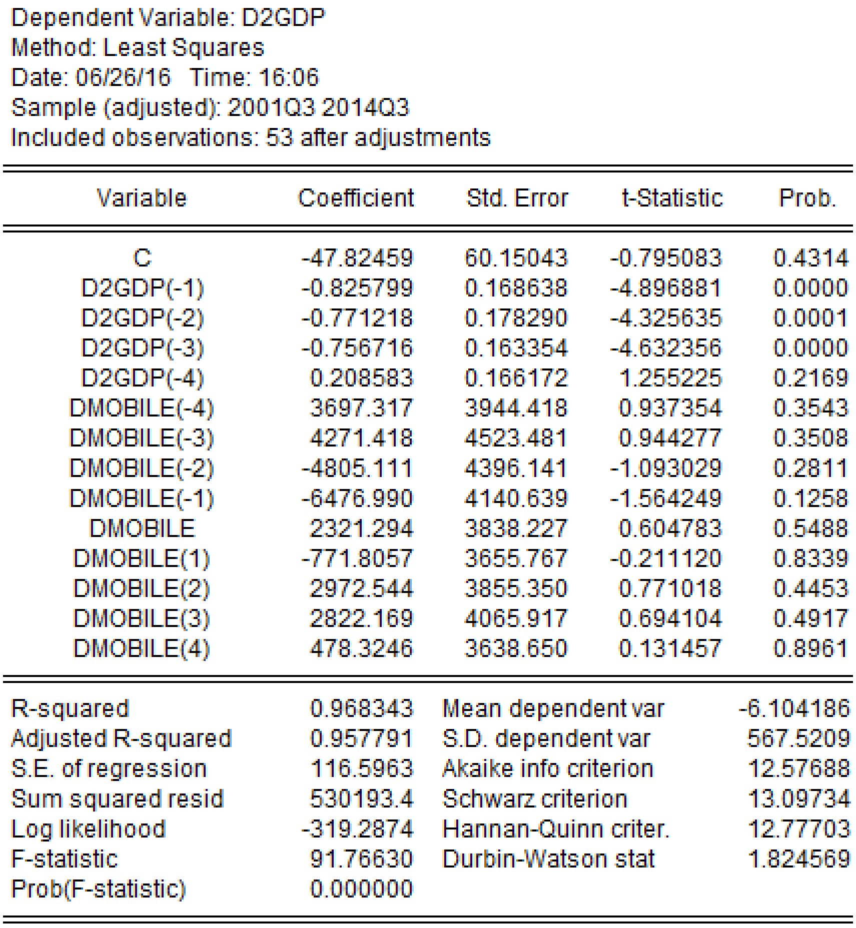


```
Wald Test:
Equation: Untitled
=====================================================
Test Statistic        Value        df       Probability
=====================================================
F-statistic         1.451793     (4, 39)      0.2355
Chi-square          5.807172        4         0.2140
=====================================================

Null Hypothesis: C(11)=C(12)=C(13)=C(14)=0
Null Hypothesis Summary:
=====================================================
Normalized Restriction (= 0)       Value       Std. Err.
=====================================================
C(11)                            -1.47E-05     1.04E-05
C(12)                            -7.77E-06     9.35E-06
C(13)                             1.58E-06     8.65E-06
C(14)                             5.43E-06     7.13E-06
=====================================================
Restrictions are linear in coefficients.
```

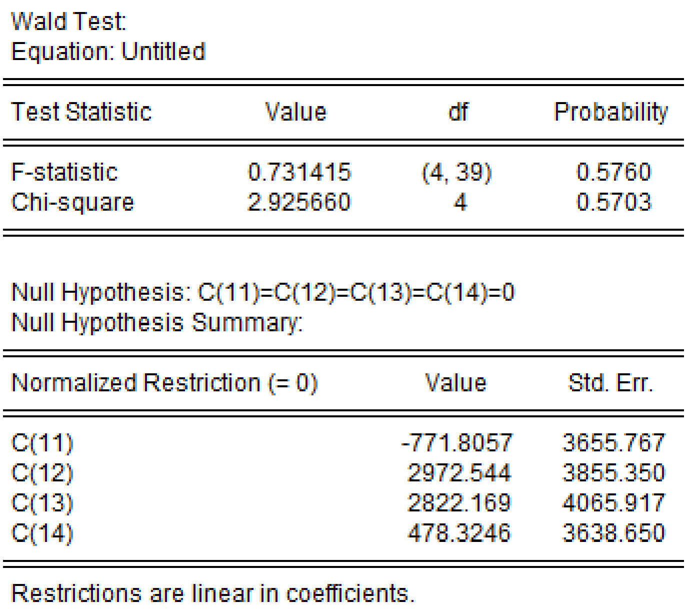
The Wald test returns a F-statistic of 1.451793 with a probability of 0.2355 (> 0.05). This means that we cannot reject the null hypothesis of no causality from Mobile penetration to GDP per capita.

Finally, we will examine if there is in fact causality running in the opposite direction (from GDP to Mobile) by running the following regression:

$$D2GDP_t = \sum_{j=-4}^{4} \theta_j DMobile_{t-j} + \sum_{i=1}^{4} \delta_i D2GDP_{t-1} + \mu + v_t$$



```
Dependent Variable: D2GDP
Method: Least Squares
Date: 06/26/16   Time: 16:06
Sample (adjusted): 2001Q3 2014Q3
Included observations: 53 after adjustments
==========================================================================
      Variable          Coefficient   Std. Error   t-Statistic    Prob.
==========================================================================
         C               -47.82459    60.15043     -0.795083     0.4314
     D2GDP(-1)           -0.825799    0.168638     -4.896881     0.0000
     D2GDP(-2)           -0.771218    0.178290     -4.325635     0.0001
     D2GDP(-3)           -0.756716    0.163354     -4.632356     0.0000
     D2GDP(-4)            0.208583    0.166172      1.255225     0.2169
    DMOBILE(-4)           3697.317    3944.418      0.937354     0.3543
    DMOBILE(-3)           4271.418    4523.481      0.944277     0.3508
    DMOBILE(-2)          -4805.111    4396.141     -1.093029     0.2811
    DMOBILE(-1)          -6476.990    4140.639     -1.564249     0.1258
      DMOBILE             2321.294    3838.227      0.604783     0.5488
    DMOBILE(1)           -771.8057    3655.767     -0.211120     0.8339
    DMOBILE(2)            2972.544    3855.350      0.771018     0.4453
    DMOBILE(3)            2822.169    4065.917      0.694104     0.4917
    DMOBILE(4)            478.3246    3638.650      0.131457     0.8961
==========================================================================
R-squared                 0.968343   Mean dependent var       -6.104186
Adjusted R-squared        0.957791   S.D. dependent var        567.5209
S.E. of regression        116.5963   Akaike info criterion     12.57688
Sum squared resid         530193.4   Schwarz criterion         13.09734
Log likelihood           -319.2874   Hannan-Quinn criter.      12.77703
F-statistic               91.76630   Durbin-Watson stat        1.824569
Prob(F-statistic)         0.000000
==========================================================================
```



Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	0.731415	(4, 39)	0.5760
Chi-square	2.925660	4	0.5703

Null Hypothesis: C(11)=C(12)=C(13)=C(14)=0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(11)	-771.8057	3655.767
C(12)	2972.544	3855.350
C(13)	2822.169	4065.917
C(14)	478.3246	3638.650

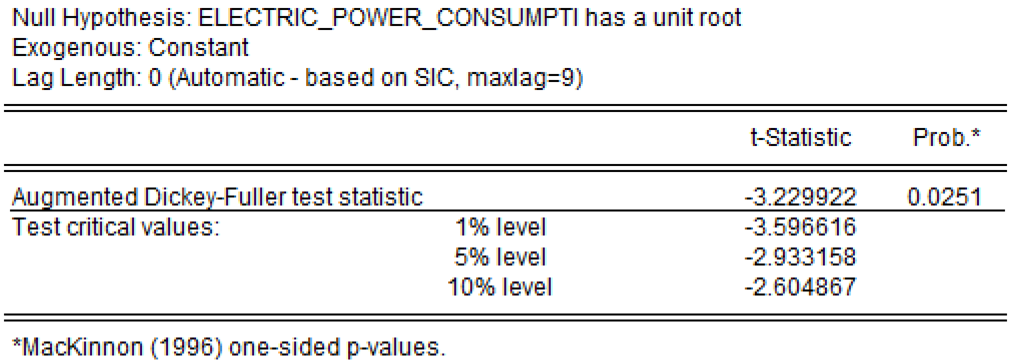
Restrictions are linear in coefficients.

We observe a F statistic of 0.731415 with a probability of 0.5760 (> 0.05), so **we can not reject the null hypothesis of no causality from GDP to Mobile penetration at the 5% significance level**.

6.2. Electricity Consumption

6.2.1. Test for Stationarity

Electric Power Consumption



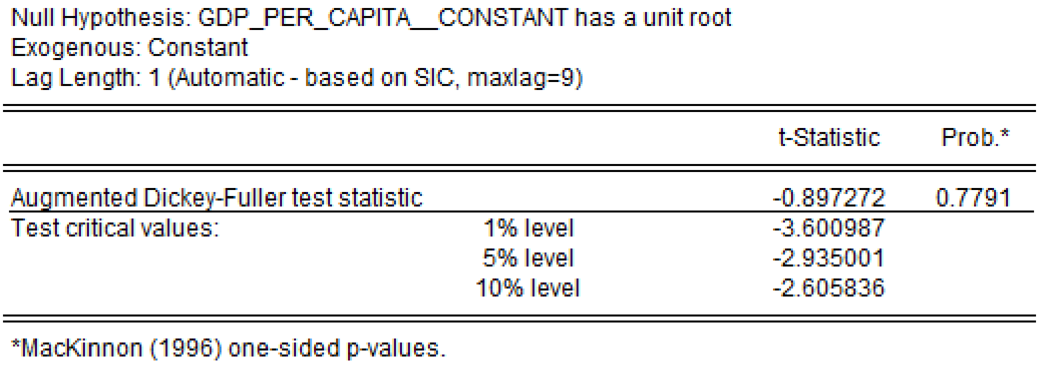
Null Hypothesis: ELECTRIC_POWER_CONSUMPTI has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=9)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.229922	0.0251
Test critical values:	1% level	-3.596616	
	5% level	-2.933158	
	10% level	-2.604867	

*MacKinnon (1996) one-sided p-values.

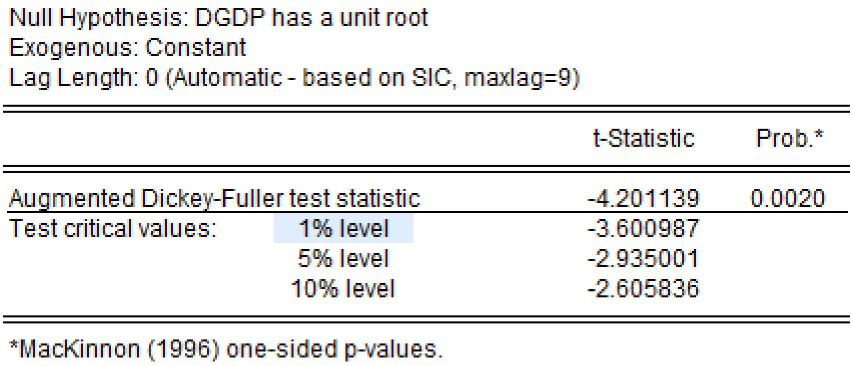
Without differencing, the null hypothesis of non-stationarity (data series has a unit root) can be rejected at the 5% significance level but not at the 1% level. As we normally use the 5% level in our analysis in this thesis, we can reject the null hypothesis that the data series is non-stationary.

GDP per capita (yearly, in constant 2005 dollars)



```
Null Hypothesis: GDP_PER_CAPITA__CONSTANT has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic - based on SIC, maxlag=9)

                                                   t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic              -0.897272     0.7791
Test critical values:            1% level           -3.600987
                                 5% level           -2.935001
                                 10% level          -2.605836

*MacKinnon (1996) one-sided p-values.
```

Unfortunately, we cannot reject the null hypothesis of non-stationarity for GDP_PER_CAPITA__CONSTANT at the 1%, 5% or even 10% level. However, after first differencing, we obtain the following results:



```
Null Hypothesis: DGDP has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=9)

                                        t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic   -4.201139     0.0020
Test critical values:      1% level      -3.600987
                           5% level      -2.935001
                           10% level     -2.605836

*MacKinnon (1996) one-sided p-values.
```

We find that the first differences of the GDP time series are stationary, i.e. we can reject the null hypothesis of non-stationarity at the 10%, 5% and even 1% significance level.

6.2.2. Modified Sims Test

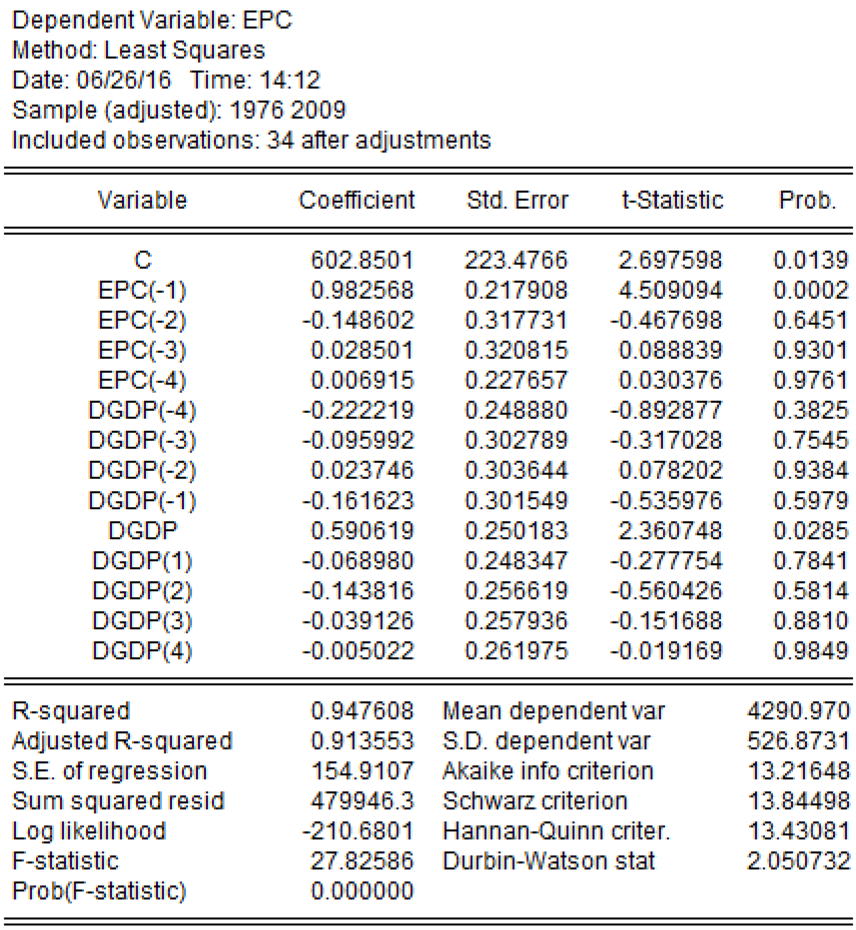
In the following section, we will use the following abbreviations:

- Electric Power Consumption (kWh per capita): **EPC**
- *1st differenced* yearly GDP Per Capita in constant 2005 US): **DGDP**

We will investigate the granger causality from EPC to DGDP for a lag period of 4 years and less.

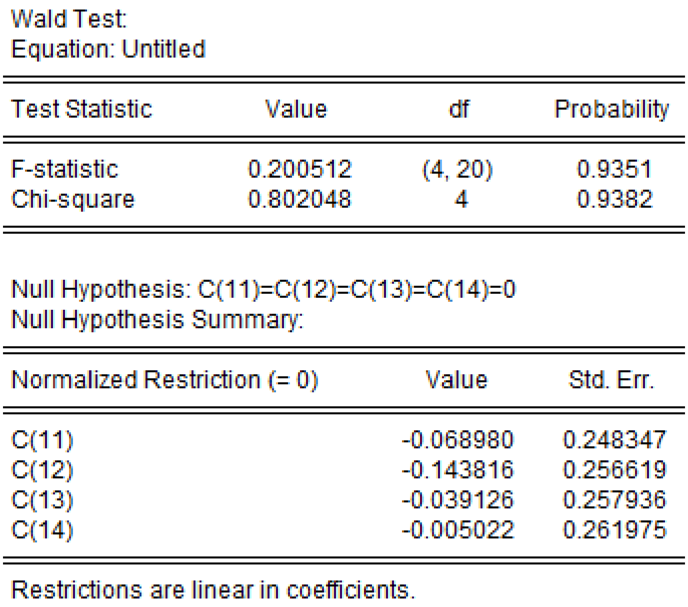
$$EPC_t = \sum_{j=-4}^{4} \theta_j DGDP_{t-j} + \sum_{i=1}^{4} \delta_i EPC_{t-1} + \mu + v_t$$

The **null hypothesis** of this test is of no (Granger) causality from EPC to DGDP. After running this regression using OLS, we get the following results:



```
Dependent Variable: EPC
Method: Least Squares
Date: 06/26/16   Time: 14:12
Sample (adjusted): 1976 2009
Included observations: 34 after adjustments
==========================================================================
      Variable        Coefficient    Std. Error    t-Statistic    Prob.
==========================================================================
          C            602.8501      223.4766      2.697598      0.0139
       EPC(-1)         0.982568      0.217908      4.509094      0.0002
       EPC(-2)        -0.148602      0.317731     -0.467698      0.6451
       EPC(-3)         0.028501      0.320815      0.088839      0.9301
       EPC(-4)         0.006915      0.227657      0.030376      0.9761
      DGDP(-4)        -0.222219      0.248880     -0.892877      0.3825
      DGDP(-3)        -0.095992      0.302789     -0.317028      0.7545
      DGDP(-2)         0.023746      0.303644      0.078202      0.9384
      DGDP(-1)        -0.161623      0.301549     -0.535976      0.5979
        DGDP           0.590619      0.250183      2.360748      0.0285
       DGDP(1)        -0.068980      0.248347     -0.277754      0.7841
       DGDP(2)        -0.143816      0.256619     -0.560426      0.5814
       DGDP(3)        -0.039126      0.257936     -0.151688      0.8810
       DGDP(4)        -0.005022      0.261975     -0.019169      0.9849
==========================================================================
R-squared                 0.947608   Mean dependent var         4290.970
Adjusted R-squared        0.913553   S.D. dependent var         526.8731
S.E. of regression        154.9107   Akaike info criterion      13.21648
Sum squared resid         479946.3   Schwarz criterion          13.84498
Log likelihood           -210.6801   Hannan-Quinn criter.       13.43081
F-statistic               27.82586   Durbin-Watson stat         2.050732
Prob(F-statistic)         0.000000
==========================================================================
```

After this test, a Wald test for coefficient restriction is run in order to obtain an F statistic and test the null hypothesis that all coefficients on the lead values of DGDP are all equal to zero.



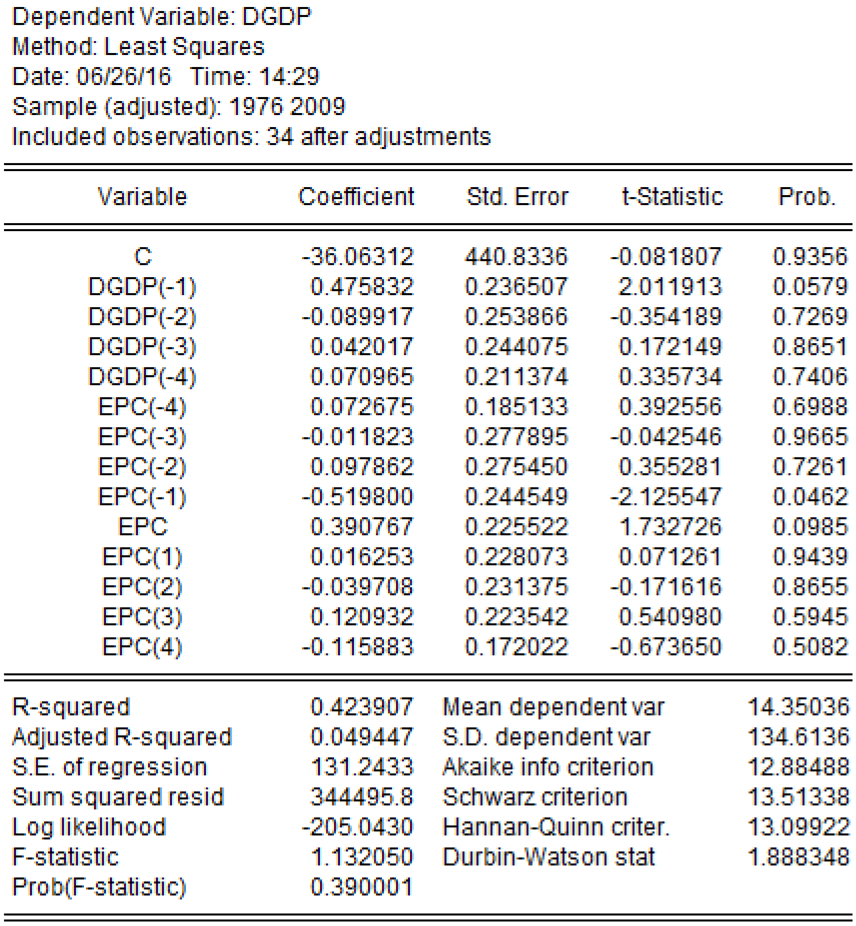
```
Wald Test:
Equation: Untitled
=======================================================
Test Statistic        Value          df       Probability
=======================================================
F-statistic          0.200512      (4, 20)      0.9351
Chi-square           0.802048        4          0.9382
=======================================================

Null Hypothesis: C(11)=C(12)=C(13)=C(14)=0
Null Hypothesis Summary:
=======================================================
Normalized Restriction (= 0)          Value       Std. Err.
=======================================================
C(11)                               -0.068980     0.248347
C(12)                               -0.143816     0.256619
C(13)                               -0.039126     0.257936
C(14)                               -0.005022     0.261975
=======================================================

Restrictions are linear in coefficients.
```

We observe a F statistic of 0.200512 with a probability of 0.9351 (> 0.05), so **we can not reject the null hypothesis of no causality from ECP to GDP at the 5% significance level**.

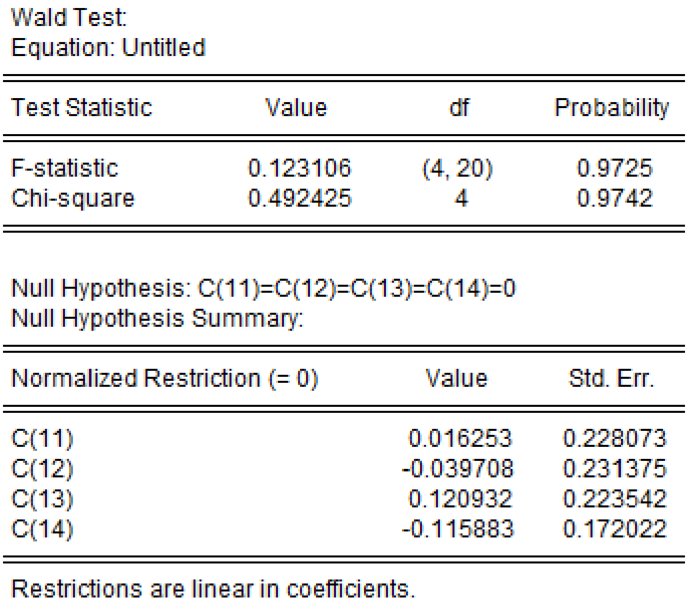
Finally, we will examine if there is in fact causality running in the opposite direction (from GDP to EPC) by running the following regression:

$$DGDP_t = \sum_{j=-4}^{4} \theta_j EPC_{t-j} + \sum_{i=1}^{4} \delta_i DGDP_{t-1} + \mu + v_t$$

This gives us the following output:



```
Dependent Variable: DGDP
Method: Least Squares
Date: 06/26/16   Time: 14:29
Sample (adjusted): 1976 2009
Included observations: 34 after adjustments
==========================================================================
        Variable         Coefficient    Std. Error    t-Statistic    Prob.
==========================================================================
           C              -36.06312     440.8336     -0.081807     0.9356
        DGDP(-1)           0.475832     0.236507      2.011913     0.0579
        DGDP(-2)          -0.089917     0.253866     -0.354189     0.7269
        DGDP(-3)           0.042017     0.244075      0.172149     0.8651
        DGDP(-4)           0.070965     0.211374      0.335734     0.7406
         EPC(-4)           0.072675     0.185133      0.392556     0.6988
         EPC(-3)          -0.011823     0.277895     -0.042546     0.9665
         EPC(-2)           0.097862     0.275450      0.355281     0.7261
         EPC(-1)          -0.519800     0.244549     -2.125547     0.0462
           EPC             0.390767     0.225522      1.732726     0.0985
         EPC(1)            0.016253     0.228073      0.071261     0.9439
         EPC(2)           -0.039708     0.231375     -0.171616     0.8655
         EPC(3)            0.120932     0.223542      0.540980     0.5945
         EPC(4)           -0.115883     0.172022     -0.673650     0.5082
==========================================================================
R-squared                  0.423907   Mean dependent var         14.35036
Adjusted R-squared         0.049447   S.D. dependent var         134.6136
S.E. of regression         131.2433   Akaike info criterion      12.88488
Sum squared resid          344495.8   Schwarz criterion          13.51338
Log likelihood            -205.0430   Hannan-Quinn criter.       13.09922
F-statistic                1.132050   Durbin-Watson stat         1.888348
Prob(F-statistic)          0.390001
==========================================================================
```



```
Wald Test:
Equation: Untitled
=====================================================================
Test Statistic              Value            df          Probability
=====================================================================
F-statistic               0.123106        (4, 20)         0.9725
Chi-square                0.492425          4             0.9742
=====================================================================

Null Hypothesis: C(11)=C(12)=C(13)=C(14)=0
Null Hypothesis Summary:
=====================================================================
Normalized Restriction (= 0)               Value           Std. Err.
=====================================================================
C(11)                                     0.016253         0.228073
C(12)                                    -0.039708         0.231375
C(13)                                     0.120932         0.223542
C(14)                                    -0.115883         0.172022
=====================================================================
Restrictions are linear in coefficients.
```

We observe a F statistic of 0.12306 with a probability of 0.9725 (> 0.05), so **we can not reject the null hypothesis of no causality from GDP to ECP at the 5% significance level**.

6.2. Female Genital Cutting (FGC)

In this part, the art and science of **machine learning** will be utilized to train a model to predict the FGC state of a girl as accurately as possible. Having a large dataset available with the circumcision state already present for a large number of African countries, this leads us to the realm of **supervised learning**. The FGC dependent variable is characterized by **discrete class labels** (0 for uncircumcised, 1 for circumcised), which is typical of a **classification** task.

The different models will help us determine what the exact effect is of **individual** (mobile telephone) and **mass** (radio and television) communication technologies on the FGC state of women in Africa. This can only be accurately examined in the presence of other relevant variables.

In the next parts, the various steps making up the preprocessing, training, tuning and validation of various models will be discussed in detail.

6.2.1. Data Preprocessing

The sample is selected based on the following conditions:

- The girl was circumcised (or estimated circumcised if she never underwent the procedure) **between the year of the interview and 5 years before**. This was done to ensure that the state of the variables (eg. possession of radio, mobile telephone etc.) recorded during the interview was already present at the time of circumcision. As a consequence, only *girl observations* (daughters of the mother interviewed) are present in the sample

- To limit the number of dummy variables, only observations with ethnicities of at least **500** members and religions of at least **5000** members were retained in the sample.

By selecting girls who were circumcised (or estimated to have have been circumcised for non –FGC individuals) in a similar time frame, **any differences in the average age at circumcision is abstracted** from the analysis. The only effect to keep in mind is that the number of observations is not perfectly equal between all countries. However, as one can see from the table below, which shows the number of observations for each country and percentage of total for that country, the analysis will not be overwhelmingly skewed towards one country in particular.

Country	N	Percentage
Benin	2736	9.69%
Burkina Faso	9793	28.00%
Central African Republic	1024	3.40%
Ghana	22	0.01%
Guinea	6594	36.77%
Ivory Coast	3215	16.95%
Mali	5367	23.87%
Mauritania	-	-
Nigeria	9836	15.33%
Senegal	3157	18.76%
Sierra Leone	-	-
Togo	2069	12.95%

Dependent variable

Female Circumcision (FCIRC): has the girl been circumcised (as reported by the mother during the time of interview)?

Independent variables

- **Religion**: to what religion does the head of the household belong to?
- **Ethnicity**: what ethnical background does the household have?
- **Urban vs Rural**: is the household located in a urban or rural location?
- **Has Electricity**: is the household connected to the electrical grid?
- **FCIRC Mother**: is the mother circumcised?
- **Has Refrigerator**: does the household possess a refrigerator?
- **Has Bicycle**: does the household possess at least one bicycle?
- **Has Moto**: does the household possess at least one moto?
- **Has Car**: does the household possess at least one car?
- **Literacy**: can the mother read? How well?
- **Has Mobile Telephone**: does the household possess at least one mobile?
- **Mass Media**: does the household possess a radio and/or a television?
- **Country**: in which country does the household live? – to have a country specific intercept

The different possession related variables were incorporated in the independent variables set in order to distinguish between different forms of economic wealth related possessions. There are three transport related possessions in increasing terms of income (**bicycle**, **moto** and **car**), two communication related ones (**mobile telephone** and **mass media**) and one (almost) purely income related possession, being a **refrigerator**.

6.2.2. Feature selection

Often, a model performs better on the training dataset than on the test dataset. This phenomenon is a sign of **overfitting**. This means that the model doesn't generalize well to unseen data because it is too complex. The additional error which appears when the model is used on new data is also called the **generalization error**.

A common way to reduce the generalization error is to:

- Collect **more data** to train the model on
- Use **regularization** (l1 or l2) to introduce a penalty for more complex models. The concept of regularization consists of adding a term to the objective function of a model as such to penalize excessive parameter weights.
- Choose a model specification with **fewer parameters**
- Reduce the dimensionality of the data by **feature selection** or **feature extraction**

In this analysis, we will employ the techniques of **regularization** and **feature selection** to avoid the issue of overfitting.

In this section related to **feature selection**, two techniques are being used: a sequential backward selection (SBS) algorithm and the importance scores obtained when running a Random Forest model.

Sequential Backward Selection (SBS)

The SBS algorithm sequentially removes features from the feature subspace until the number of desired features is reached. The algorithm selects the feature to remove by examining a **criterion function J**, this is typically the performance before and after removal of the feature. More concretely, at each stage the feature is removed that causes the lowest decrease in performance when removed.

In our algorithm, **accuracy** (*the percentage of correctly classified observations)* is used in the criterion function. The model used to implement the SBS algorithm is a **logistic regression.** All variables which cause the increase from zero to the maximum accuracy are retained as relevant variables.

In order to avoid the obvious data mining pitfalls in using the same data partition to build and validate our model on, the training data fed to the SBS algorithm is further split into a training and test set.

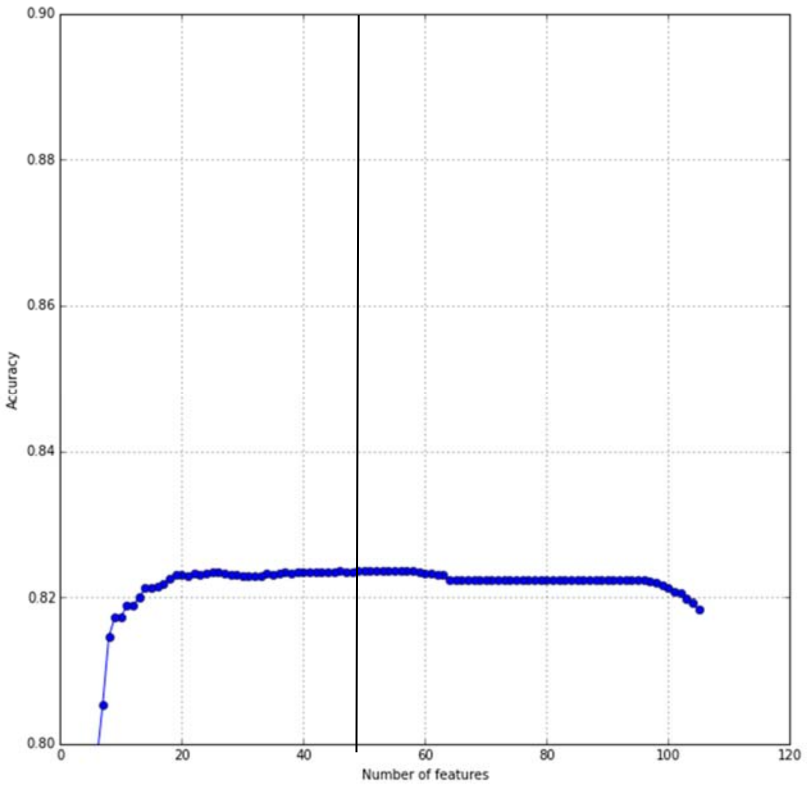
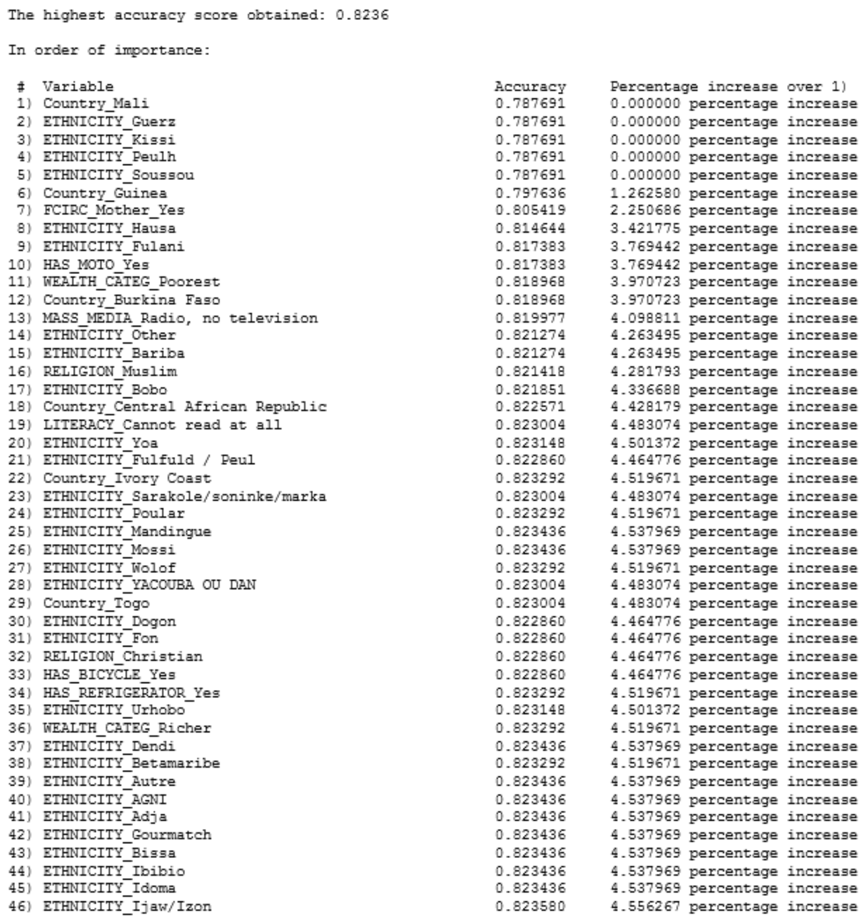


Figure 12: Accuracy of the SBS algorithm



```
The highest accuracy score obtained: 0.8236

In order of importance:

 #  Variable                                          Accuracy      Percentage increase over 1)
 1) Country_Mali                                      0.787691      0.000000 percentage increase
 2) ETHNICITY_Guerz                                   0.787691      0.000000 percentage increase
 3) ETHNICITY_Kissi                                   0.787691      0.000000 percentage increase
 4) ETHNICITY_Peulh                                   0.787691      0.000000 percentage increase
 5) ETHNICITY_Soussou                                 0.787691      0.000000 percentage increase
 6) Country_Guinea                                    0.797636      1.262580 percentage increase
 7) FCIRC_Mother_Yes                                  0.805419      2.250686 percentage increase
 8) ETHNICITY_Hausa                                   0.814644      3.421775 percentage increase
 9) ETHNICITY_Fulani                                  0.817383      3.769442 percentage increase
10) HAS_MOTO_Yes                                      0.817383      3.769442 percentage increase
11) WEALTH_CATEG_Poorest                              0.818968      3.970723 percentage increase
12) Country_Burkina Faso                              0.818968      3.970723 percentage increase
13) MASS_MEDIA_Radio, no television                   0.819977      4.098811 percentage increase
14) ETHNICITY_Other                                   0.821274      4.263495 percentage increase
15) ETHNICITY_Bariba                                  0.821274      4.263495 percentage increase
16) RELIGION_Muslim                                   0.821418      4.281793 percentage increase
17) ETHNICITY_Bobo                                    0.821851      4.336688 percentage increase
18) Country_Central African Republic                  0.822571      4.428179 percentage increase
19) LITERACY_Cannot read at all                       0.823004      4.483074 percentage increase
20) ETHNICITY_Yoa                                     0.823148      4.501372 percentage increase
21) ETHNICITY_Fulfuld / Peul                          0.822860      4.464776 percentage increase
22) Country_Ivory Coast                               0.823292      4.519671 percentage increase
23) ETHNICITY_Sarakole/soninke/marka                  0.823004      4.483074 percentage increase
24) ETHNICITY_Poular                                  0.823292      4.519671 percentage increase
25) ETHNICITY_Mandingue                               0.823436      4.537969 percentage increase
26) ETHNICITY_Mossi                                   0.823436      4.537969 percentage increase
27) ETHNICITY_Wolof                                   0.823292      4.519671 percentage increase
28) ETHNICITY_YACOUBA OU DAN                          0.823004      4.483074 percentage increase
29) Country_Togo                                      0.823004      4.483074 percentage increase
30) ETHNICITY_Dogon                                   0.822860      4.464776 percentage increase
31) ETHNICITY_Fon                                     0.822860      4.464776 percentage increase
32) RELIGION_Christian                                0.822860      4.464776 percentage increase
33) HAS_BICYCLE_Yes                                   0.822860      4.464776 percentage increase
34) HAS_REFRIGERATOR_Yes                              0.823292      4.519671 percentage increase
35) ETHNICITY_Urhobo                                  0.823148      4.501372 percentage increase
36) WEALTH_CATEG_Richer                               0.823292      4.519671 percentage increase
37) ETHNICITY_Dendi                                   0.823436      4.537969 percentage increase
38) ETHNICITY_Betamaribe                              0.823292      4.519671 percentage increase
39) ETHNICITY_Autre                                   0.823436      4.537969 percentage increase
40) ETHNICITY_AGNI                                    0.823436      4.537969 percentage increase
41) ETHNICITY_Adja                                    0.823436      4.537969 percentage increase
42) ETHNICITY_Gourmatch                               0.823436      4.537969 percentage increase
43) ETHNICITY_Bissa                                   0.823436      4.537969 percentage increase
44) ETHNICITY_Ibibio                                  0.823436      4.537969 percentage increase
45) ETHNICITY_Idoma                                   0.823436      4.537969 percentage increase
46) ETHNICITY_Ijaw/Izon                               0.823580      4.556267 percentage increase
```

Random Forest Importances

A Random Forests model calculates feature importance as **the average impurity decrease** computed from all decision trees in the forest. This technique doesn't make any assumptions if the data is linearly separable or not. This model will be more discussed later on.

All variables (sorted by importance) which cumulatively reach 90% of importance are selected as relevant variables

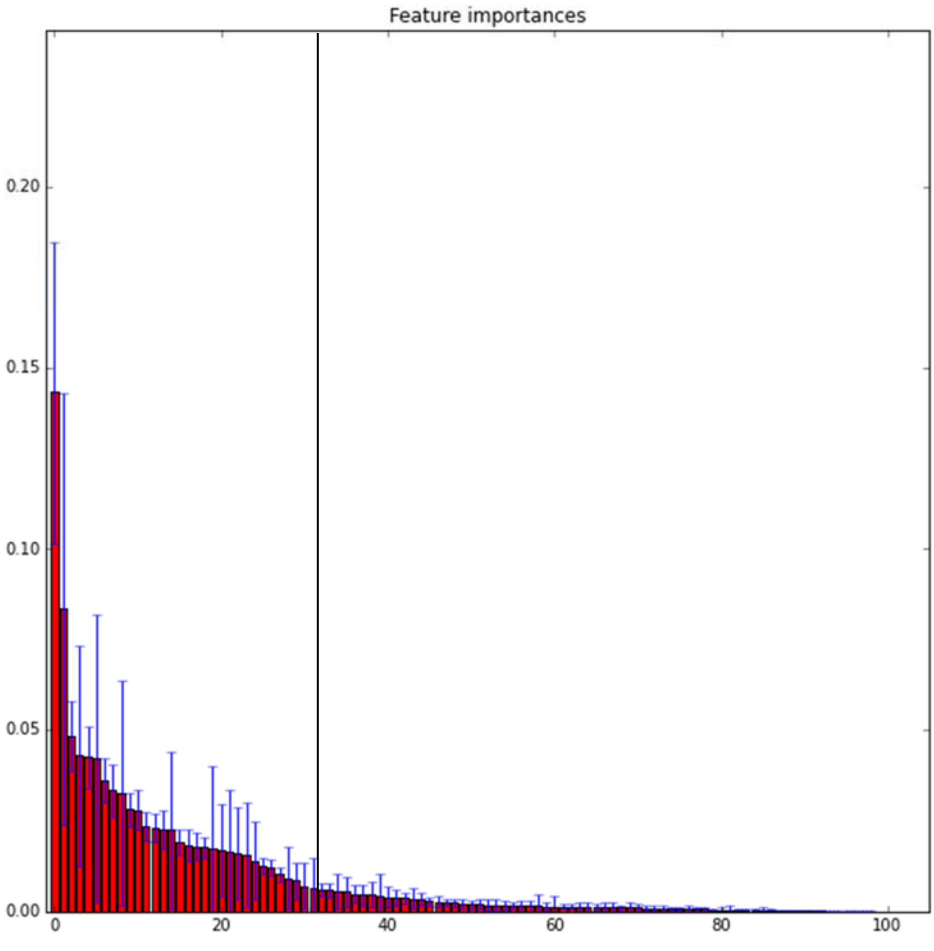
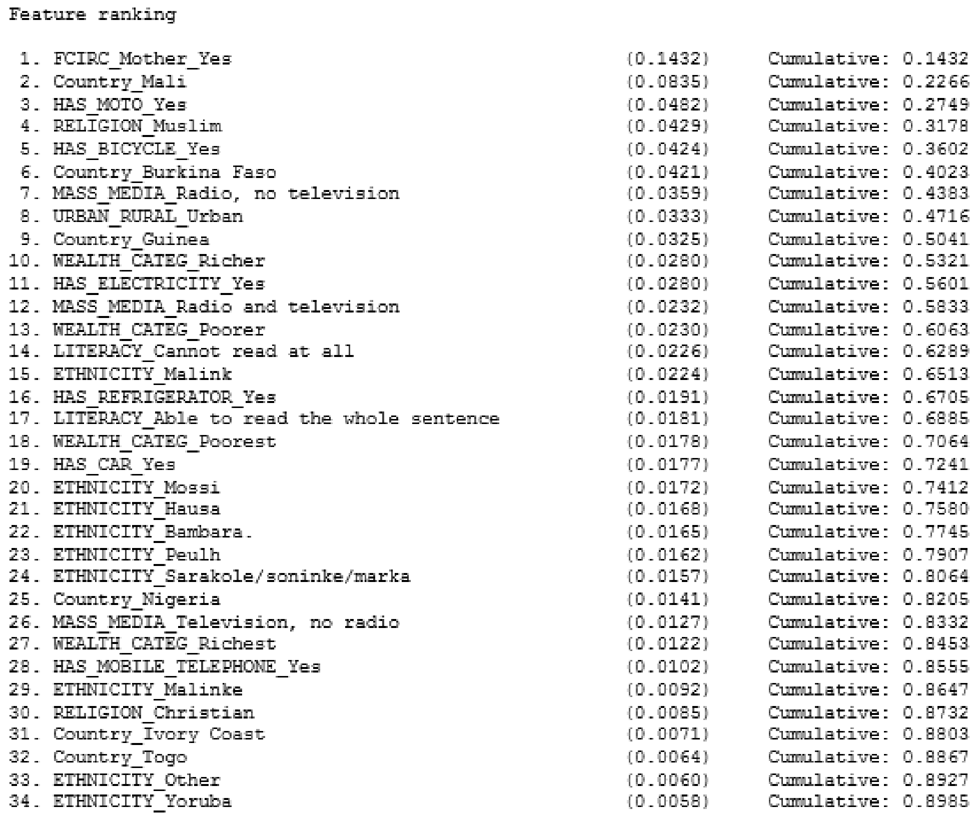


Figure 13: Average impurity decrease per feature for the Random Forest algorithm

As can be seen from the table below, the first 10 variables together make up for more than 50% of the cumulative importance scores.

Variable	Importance score	Cumulative importance
FCIRC Mother	14.32%	14.32%
Country Mali	8.35%	22.66%
Has Moto	4.82%	27.49%
Religion Muslim	4.29%	31.78%
Has Bicycle	4.24%	36.02%
Country Burkina Faso	4.21%	40.23%
Radio, no television	3.59%	43.83%
Urban	3.33%	47.16%

Country Guinea	3.25%	50.41%
Wealth - Richer	2.80%	53.21%



```
Feature ranking

 1. FCIRC_Mother_Yes                                  (0.1432)    Cumulative: 0.1432
 2. Country_Mali                                      (0.0835)    Cumulative: 0.2266
 3. HAS_MOTO_Yes                                      (0.0482)    Cumulative: 0.2749
 4. RELIGION_Muslim                                   (0.0429)    Cumulative: 0.3178
 5. HAS_BICYCLE_Yes                                   (0.0424)    Cumulative: 0.3602
 6. Country_Burkina Faso                              (0.0421)    Cumulative: 0.4023
 7. MASS_MEDIA_Radio, no television                   (0.0359)    Cumulative: 0.4383
 8. URBAN_RURAL_Urban                                 (0.0333)    Cumulative: 0.4716
 9. Country_Guinea                                    (0.0325)    Cumulative: 0.5041
10. WEALTH_CATEG_Richer                               (0.0280)    Cumulative: 0.5321
11. HAS_ELECTRICITY_Yes                               (0.0280)    Cumulative: 0.5601
12. MASS_MEDIA_Radio and television                   (0.0232)    Cumulative: 0.5833
13. WEALTH_CATEG_Poorer                               (0.0230)    Cumulative: 0.6063
14. LITERACY_Cannot read at all                       (0.0226)    Cumulative: 0.6289
15. ETHNICITY_Malink                                  (0.0224)    Cumulative: 0.6513
16. HAS_REFRIGERATOR_Yes                              (0.0191)    Cumulative: 0.6705
17. LITERACY_Able to read the whole sentence          (0.0181)    Cumulative: 0.6885
18. WEALTH_CATEG_Poorest                              (0.0178)    Cumulative: 0.7064
19. HAS_CAR_Yes                                       (0.0177)    Cumulative: 0.7241
20. ETHNICITY_Mossi                                   (0.0172)    Cumulative: 0.7412
21. ETHNICITY_Hausa                                   (0.0168)    Cumulative: 0.7580
22. ETHNICITY_Bambara.                                (0.0165)    Cumulative: 0.7745
23. ETHNICITY_Peulh                                   (0.0162)    Cumulative: 0.7907
24. ETHNICITY_Sarakole/soninke/marka                  (0.0157)    Cumulative: 0.8064
25. Country_Nigeria                                   (0.0141)    Cumulative: 0.8205
26. MASS_MEDIA_Television, no radio                   (0.0127)    Cumulative: 0.8332
27. WEALTH_CATEG_Richest                              (0.0122)    Cumulative: 0.8453
28. HAS_MOBILE_TELEPHONE_Yes                          (0.0102)    Cumulative: 0.8555
29. ETHNICITY_Malinke                                 (0.0092)    Cumulative: 0.8647
30. RELIGION_Christian                                (0.0085)    Cumulative: 0.8732
31. Country_Ivory Coast                               (0.0071)    Cumulative: 0.8803
32. Country_Togo                                      (0.0064)    Cumulative: 0.8867
33. ETHNICITY_Other                                   (0.0060)    Cumulative: 0.8927
34. ETHNICITY_Yoruba                                  (0.0058)    Cumulative: 0.8985
```

6.2.3. Training of different models

The following models will be trained and compared:

- Logistic Regression
- Random Forest
- K Nearest Neighbours (KNN)

Ensemble methods: combining different models to obtain a better performance than its individual members

- Majority Vote
- Bagging
- Adaptive Boosting

The first model (logistic regression) will be used to illustrate all different steps and methods of the analysis in depth. Afterwards, the results of the analysis for all other methods will be compared in a more sumeer fashion.

Logistic Regression

This model is very easy to implement **but** is most performing when the underlying data is **linearly separable**.

A key aspect of a logistic regression model is the **odds ratio**, which is expressed as:

$$OR = \frac{p}{(1-p)}$$

In which ‘p’ is the probability of presence of the relevant feature (female circumcision in the current analysis) and as a consequence ‘p-1’ is the probability of the absence of that feature (not circumcised).

Tuning the parameters using Grid Search

In order to maximize the **accuracy** of classification, an exhaustive (grid) search is undertaken to determine the optimal set of model parameters. The following parameter values are searched:

- **C** (the inverse of lambda, the regularization coefficient): [0.01, 0.1, 1.0, 10.0]
- **Regularization method**: l1 or l2
- **Intercept in the model**? True or False
- **Solver**: newton-cg, liblinear or lbfgs

The output obtained:

```
The best accuracy score obtained is 0.8097
The model with the highest accuracy has the following parameters:
{'penalty': 'l1', 'C': 0.1, 'solver': 'liblinear', 'fit_intercept': True}
Test accuracy: 0.8112
```

This means that the best performing model is capable of classifying women correctly in 80% of cases

Odds ratio

Variable	Odds ratio
FCIRC_Mother_Yes	17.32222336
Country_Mali	4.600915051
ETHNICITY_Hausa	3.527077756
Country_Nigeria	2.80887944
Country_Guinea	2.573505695
Country_Central African Republic	2.457456119
ETHNICITY_Fulani	2.300111609
ETHNICITY_Poular	2.257081414
ETHNICITY_Sarakole/soninke/marka	2.028346622
ETHNICITY_Mandingue	2.01018766
RELIGION_Muslim	1.705814469

ETHNICITY_Malinke	1.702209439
ETHNICITY_Malink	1.557118492
ETHNICITY_Fulfuld / Peul	1.433282783
ETHNICITY_Yoruba	1.179271119
LITERACY_Cannot read at all	1.15747382
HAS_MOTO_Yes	1.129660771
WEALTH_CATEG_Poorest	1.093555573
ETHNICITY_Peulh	1.061239901
WEALTH_CATEG_Poorer	1.054764714
HAS_MOBILE_TELEPHONE_Yes	1.046789311
MASS_MEDIA_Radio, no television	1.031998991
MASS_MEDIA_Radio and television	1.027009051
HAS_BICYCLE_Yes	1.019580987

WEALTH_CATEG_Richer	**0.980885821**
WEALTH_CATEG_Richest	0.976782043
RELIGION_Christian	0.924506583
URBAN_RURAL_Urban	0.909052065
ETHNICITY_Bissa	0.89054898
HAS_REFRIGERATOR_Yes	0.883433321
LITERACY_Able to read the whole sentence	0.864341348
ETHNICITY_Wolof	0.813965186
ETHNICITY_Kissi	0.782306461
ETHNICITY_Soussou	0.744014922
ETHNICITY_Guerz	0.658735971
ETHNICITY_Mossi	0.624408263
ETHNICITY_Yoa	0.623486543
Country_Burkina Faso	0.48769075
ETHNICITY_Ijaw/Izon	0.4183025
ETHNICITY_Urhobo	0.387934012
Country_Togo	0.377631164
ETHNICITY_Bariba	0.218221439

Learning Curve

A learning curve plots the **training** and **validation** accuracy against the size of the training set. As such, it is easy to detect two common model issues:

- If the training and validation accuracies are quite low, the model is said to have **high bias** or it **underfits** the underlying data. In short, this means that the model is not complex enough. Two common solutions are either increasing the number of parameters in the model or decreasing the degree of regularization (if present).

- If there exists a large gap between the training and validation accuracy, the model is said to have **high variance** or **overfits** the model. Common solutions are to collect more data or decrease the complexity of the model by either increasing the regularization parameter or decreasing the number of features by feature selection or feature extraction.

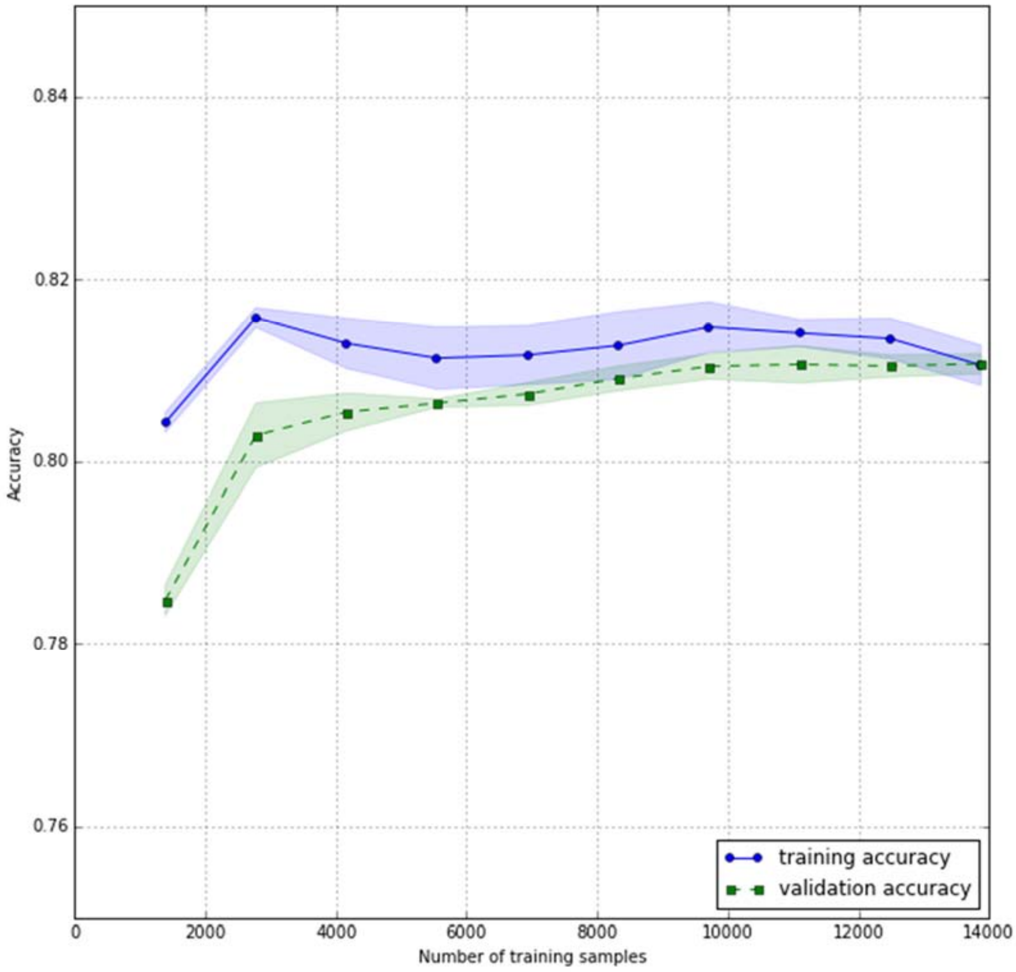


Figure 14: Learning curve

Because of the feature selection stage and the application of l1 regularization, the resulting model has low bias (the accuracy score of 85% is relatively high) and low variance (almost no overfitting: there is no large gap between training and testing accuracy).

Validation Curve (C parameter)

Validation curves are similar to learning curve as they allow to detect an address issues of over- and underfitting. However, instead of plotting the training and validation accuracy as functions of the training set size, they use one of the parameters of the model as the X-axis variable.

While the parameters of the model have already been optimized in the grid search stage of the analysis, it is worthwhile to plot the model performance for differing ranges of the model parameters to detect over- or underfitting issues for specific values.

For the logistic regression model, values of the parameter C (the inverse of the regularization coefficient lambda) are plotted on the X axis.

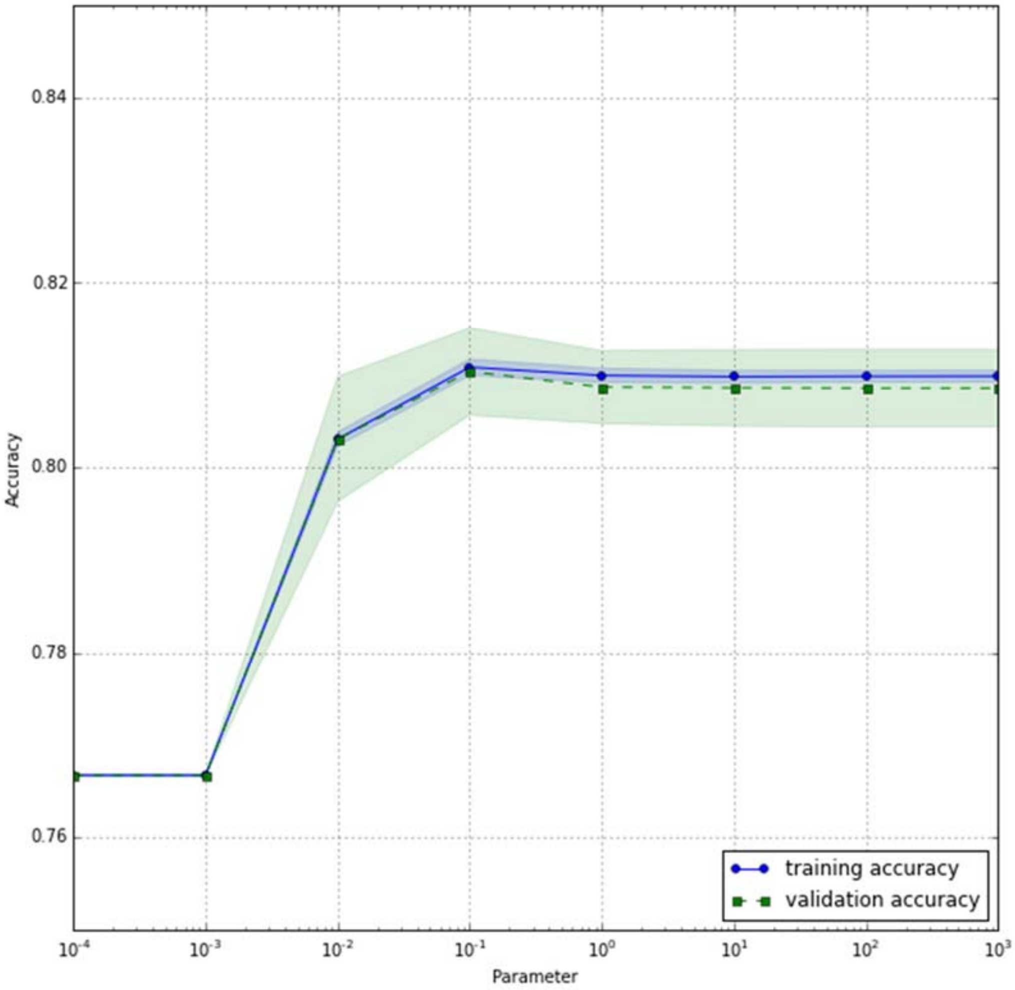


Figure 15: Validation curve

One can clearly see that there are no issues of overfitting present for each of the values of C. However, for very low values (which translated to strong regularization) the model underfits the data significantly. As already found in the Grid search optimization, the optimum value for C is clearly 10^{-1} or 0.1.

Receiver Operator Characteristic (ROC)

ROC curves plot the **true positive rate** versus the **false positive rate**. The exists a trade-off between these rates for different decision thresholds of the classifier. As can be seen from the graph, the diagonal line corresponds with a purely 'random guessing' strategy while the perfect classifier would fall in the top left corner of the graph.

The **area under the curve (AUC)** functions as another performance indicator with a maximum value of 1 and a minimum (corresponding to random guessing) value of 0.5.

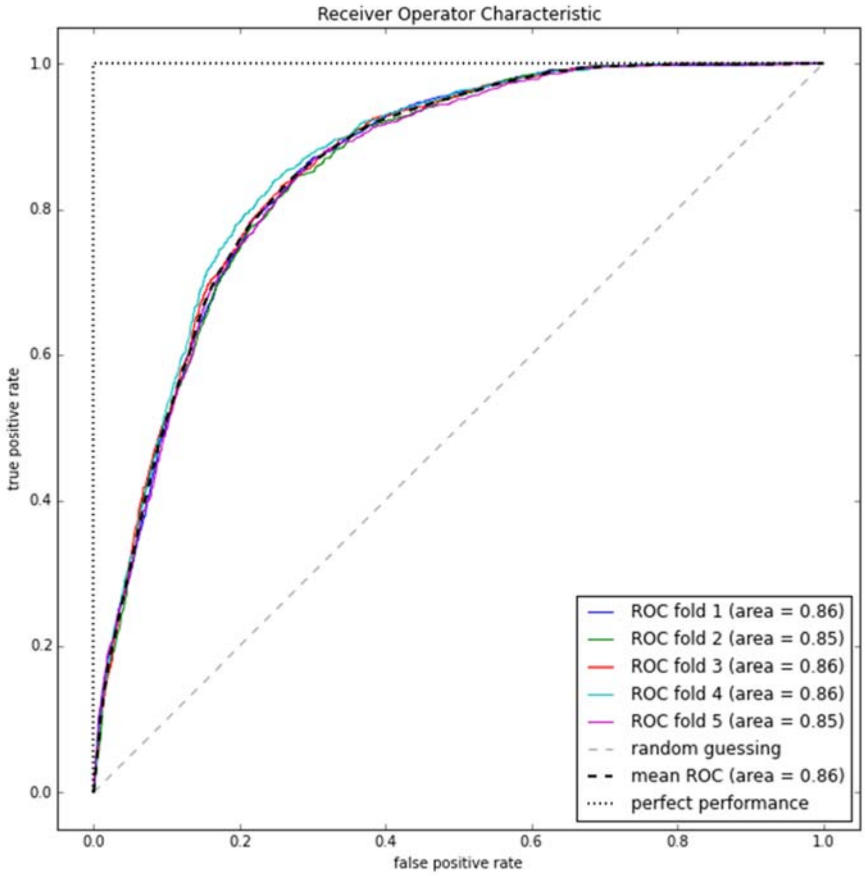
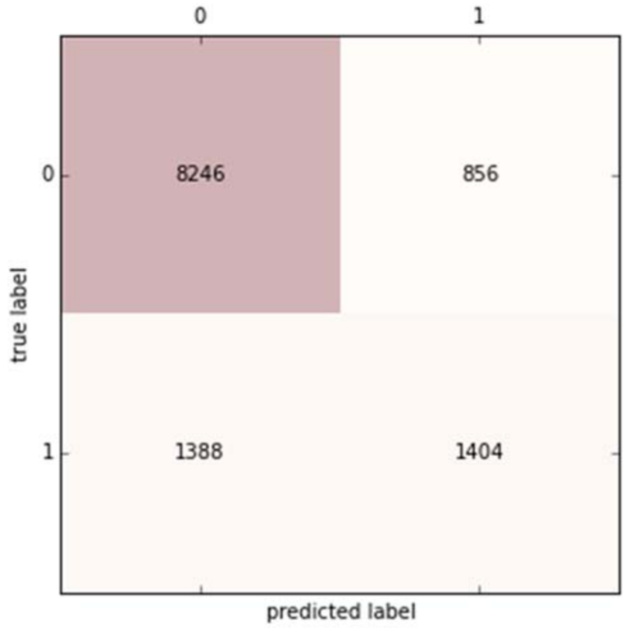


Figure 16: Receiver Operator Characteristic (ROC)

Other Performance Evaluation Metrics



Error score: 0.1887

Accuracy score: 0.8113

Precision: 0.6212

Recall: 0.5029

F1: 0.5558

Random Forest

Tuning the parameters using Grid Search

```
The best accuracy score obtained is 0.8129
The  model  with  the  highest  accuracy  has  the  following  parameters:
{'max_features': 'sqrt', 'criterion': 'gini', 'max_depth': 12}
Test accuracy: 0.8171
```

Performance indicators

Performance indicator	
Accuracy	81.71%
Error score	18.42%
Precision	63.27%
Recall	51.33%
F1	56.67%
AUC	0.86

Support Vector Machine (SVM)

Tuning the parameters using Grid Search

```
The best accuracy score obtained is 0.8114
The model with the highest accuracy has the following parameters:
{'kernel': 'poly', 'C': 10.0, 'degree': 2}
Test accuracy: 0.8129
```

Performance indicators

Performance indicator	
Accuracy	81.29%
AUC	0.82

K-Nearest Neighbors (KNN)

Tuning the parameters using Grid Search

```
The best accuracy score obtained is 0.7891
The model with the highest accuracy has the following parameters:
{'n_neighbors': 4, 'metric': 'minkowski', 'p': 1}
Test accuracy: 0.7947
```

Performance indicators

Performance indicator	
Accuracy	79.47%
AUC	0.80

Bernoulli Naive Bayes

Tuning the parameters using Grid Search

```
The best accuracy score obtained is 0.7861
The model with the highest accuracy has the following parameters: {'alpha':
1, 'fit_prior': True}
Test accuracy: 0.7823
```

Performance indicators

Performance indicator	
Accuracy	78.23%
AUC	0.83

7. Conclusions

At the macro level, the example of South Africa was used to detect the direction of causality between GDP per capita and respectively electricity consumption and mobile penetration. A modified Sims test was applied to the data but no unilateral or bilateral causality could be detected. The reasons for this could be that our sample was not large enough to obtain a significant result. In future research, an elaborate panel study econometric model could be researched, but this is out of the scope of this thesis.

At the micro level, techniques of machine learning were applied to the USAID and UNICEF data for a large set of African countries. While a lot of determinants were identified, the hypothesis that increased mobile penetration could have a measurable impact of the abandonment of FGC, could not be confirmed. The main determinants found were (in order of magnitude): FGC status of the mother, ethnicity, religion and wealth.

References

WHO 2014: http://www.who.int/mediacentre/factsheets/fs241/en/

Adinma, J., 1997. Current status of female circumcision among Nigerian Igbos. WAJM 16, 227-231.

Aker, J., 2008. Does Digital Divide or Provide. The Impact of Cell Phones on Grain Markets in Niger. BREAD Working Paper No. 177, February 2008.

Arellano M. & Bond S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, Review of Economic Studies 58, 277–297.

Aschauer, David A, 1989. Highway capacity and economic growth. Economic Perspectives, 14-24.

Bayes, A., 2001. Infrastructure and rural development: insights from a Grameen Bank village phone initiative in Bangladesh. Agricultural Economics 25 (2001) 261-272.

Bebee, F. L., and E. J. W. Gilling, 1976. Telecommunications and Economic Development: a Model for Planning and PolicyMaking, Telecommunications Journal, 43(7), pp. 537-543.

Chakraborty, C., & Nandi, B., 2003. Privatization, telecommunications and growth in selected Asian countries: An econometric analysis. Communications and Strategies, 52(4), 31-47.

Chong, A., Galdo, V., and Torero, M., 2006. Does Privatization Deliver? Access to Telephone Services and Household Income in Poor Rural Areas Using a Quasi-Natural Experiment for Peru. Contributed paper for the International Association of Agricultural Economists Conference, August 2006.

Cieślik A., Kaniewska M., 2004. Telecommunications Infrastructure and Regional Economic Development: The Case of Poland, Regional Studies, 38:6, 713-725, DOI: 10.1080/003434042000240996

Cronin, J. F., Parker, E. B., Colleran, E. K. & Gold, M. A., 1991. Telecommunications infrastructure and economic growth: an analysis of causality, Telecommunications Policy, 15, pp. 529–535.

Cronin, J. F., Coleran, E. K., Herbert, P. L. & Lewitzky, S., 1993. Telecommunications and growth: the contribution of telecommunication infrastructure investment to aggregate and sectional productivity, Telecommunications Policy, August, pp. 415–430.

Datta A., Agarwal S., 2004. Telecommunications and economic growth: a panel data approach, Applied Economics, 36:15, 1649-1654, DOI: 10.1080/0003684042000218552

DeLong, J., Summers, L., 1991. Equipment investment and economic growth. Q. J. Econ. 106, 445-502.

Dhokalia, R. R. and B. Harlam, 1994. Telecommunications and Economic Development, Telecommunications Policy, vol. 18 (6), pp. 470-477.

Duncombe, R., 2012. Understanding Mobile Phone Impact on Livelihoods in Developing Countries: A New Research Framework. Development Informatics Working Paper No 48, Institute for Development Policy and Management, The University of Manchester.

Economides, N., Himmelberg, C., 1995. Critical Mass and Network Size with Application to the US Fax Market. Economics Working Papers, Stern School, NYU, EC-95-11.

Egyption Fertility Care Society (EFCS), 1996. Clinic-Based Investigation of the Typology and Self-Reporting of FGM in Egypt. EFCS, Cairo.

Ellis, F., Bahiigwa, G., 2003. Livelihoods and rural poverty reduction in Uganda, World Development, 31(6): 997-1013.

Ellis, F., 2000. Rural Livelihoods and Diversity in Developing Countries, Oxford University Press, Oxford.

Fischer, S., Sahay, R., Vegh, C.A., 1996a. Economies in transition: The beginnings of growth. Papers and Proceedings of the 108th Annual Meeting of the American Economic Association, San Francisco, CA, January 5–7, 1996.

Fischer, S., Sahay, R., Vegh, C.A., 1996b. Stabilisation and growth in transition economies: The early experience. J. Econ. Perspectives 10, 45–66.

Garbade, K. D., Silber, W.L., 1978. Technology, communication and the performance of financial markets: 1840-1975. Journal of Finance, 33 (3): 819-831.

Granger, C. W. J., (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". Econometrica 37 (3): 424–438.

Greenstein, S., Spiller P.,, 1996. Estimating the Welfare Effects of Digital Infrastructure, NBER Working Papers 5770, National Bureau of Economic Research, Inc.

Gruber and Koutroumpis

Hardy, A. P., 1980. The role of the telephone in economic development, Telecommunications Policy, December, 4, pp. 278–286.

Heeks, R., 1999. Information and Communication Technologies, Poverty and Development. Development Informatics Working Paper No 5, Institute for Development Policy and Management, The University of Manchester.

International Telecommunications Union (ITU). 1998. World Telecommunication Development Report: Universal Access. Geneva, Switzerland: ITU.

Jackson, E.E., P. Akweongo, E. Sakeah, A. Hodgson, R. Asuru, and J.E. Phillips. 2003. “Inconsistent Reporting of Female Genital Cutting Status in Northern Ghana: Explanatory Factors and Analytical Consequences.” Studies in Family Planning 34(3): 200-210.

Jensen, R. (2007). The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector. The Quarterly Journal of Economics August 2007, Vol. 122, Issue3, 879-924.

Jensen, R. and Oster, E. (2009). The Power of TV: Cable Television and Women's Status in India. The Quarterly Journal of Economics (2009) 124 (3): 1057-1094.

Jipp, A. (1963). Wealth of nations and telephone density, Telecommunications Journal 30, 199-201.

Klouman, E., R. Manongi, and K.I. Klepp. 2005. “Self-reported and Observed Female Genital Cutting in Rural Tanzania: Associated Demographic Factors, HIV and Sexually Transmitted Infections.” Tropical Medicine and International Health 10(1): 105-115.

Leff, N. H., 1984. Externalities, Information costs, and Social Benefit-Cost Analysis for Economic Development: An Example from Telecommunication. Economic Development and Cultural Change, 32(2):255-276.

Levine, R., Renelt, D., 1992. A sensitivity analysis of cross-country growth regressions. Am. Econom. Rev. 82, 942-963.

Lichtenberg, F. R. (1995), ‘The Output Contribution of Computer Equipment and Personnel: A Firm-Level Analysis’, Economics of Innovation and New Technology, 3, pp. 201-217.

Madden, G. and Scott J. Savage, 1998. CEE Telecommunications Investment and Economic Growth, Information Economics and Policy, V10 (2), pp. 173-195.

Matambalya, F., and S. Wolf, 2001. The Role of ICT for the Performance of SMEs in East Africa: Empirical Evidence from Kenya and Tanzania. ZEF Discussion Papers on Development Policy 42 . Bonn, Germany: Center for Development Research (ZEF).

Muto, Megumi. (2008) “Impacts of mobile phone coverage expansion on market participation: panel data evidence from Uganda”. World Development, 37(12) 18871896.

North, D. C. (1995). The New Institutional Economics and Third World Development. In J. Harriss, Hunter, Janet & Colin M. Lewis (Ed.), The New Institutional Economics and Third World Development. London and New York: Routledge.

Norton, S.W., 1992. Transaction Costs, Telecommunications and the Microeconomics of Macroeconomic Growth, Economic Development and Cultural Change, V 41(1), pp. 175-196

Obermeyer C (1999) Female Genital surgeries: the known, the unknown, and the unknowable. Medical Anthropology Quarterly 13, 79±106.

Odujinrin et al., 1989. A study on female circumcision in Nigeria. West African Journal of Medicine. 8(3):183-192.

Overa, R. 2006. Networks, distance and trust: telecommunications development and changing trading practices in Ghana. World Development Vol.34, No.7 (2006) 1301-1315.

Röller L. and Waverman L., 1996. Telecommunications infrastructure and economic development: A simultaneous approach, WZB Discussion Paper FS IV 96-16, Wissenschaftszentrum, Berlin.

Röller L. and Waverman L., 2001. Telecommunications infrastructure and economic development: A simultaneous approach, American Economic Review 91, 909-923.

Sala-I-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller. 2004. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." American Economic Review, 94(4): 813-835.

Saunders, R., J. Wardford, and B. Wellenius. 1994. Telecommunications and Economic Development. Baltimore, United States: The Johns Hopkins University Press.

Shinjo, K., Zhang, X., 2004. ICT capital investment and productivity growth: Granger causality in Japanese and the USA Industries.

Sridhar, K. S., Sridhar, V., 2007. Telecommunications infrastructure and economic growth: Evidence from developing countries. Applied Econometrics and International Development, 7(2).

Stigler, G, 1961. The Economics of Information. Journal of Political Economy, 69 (3), 213-225.

Tolero, M., Chowdhury, S., and Galdo, V., 2003. Willingness to pay for the rural telephone service in Bangladesh and Peru. Information Economics and Policy 15 (2003) 327-361.

Yoo, S. H., & Kwak, S. J., 2004. Information technology and economic development in Korea: a causality study. International Journal of Technology Management, 27(1), 57-67.

USAID, 2003. Guide to DHS Statistics, DHS Questionnaires and Manuals.