Reproducible Contributions in Development Economics

Bastiaan Quast

16 June 2016

Contents

Introduction **3**

1 Grandfathers and Grandsons: Effects on a Male-Only Pension Change **5**
1.1 Introduction . . . 6
1.2 Data . . . 8
1.3 Empirical methodology . . . 11
1.3.1 Identification strategy . . . 12
1.3.2 Regression specification . . . 12
1.4 Results . . . 13
1.5 Conclusions and limitations . . . 16
1.A Original Estimates . . . 22
1.B Software . . . 29

2 Making the Next Billion Demand Access **30**
2.1 Introduction . . . 30
2.2 Data . . . 34
2.3 Empirical Methodology . . . 36
2.3.1 Identification Strategy . . . 36
2.3.2 Regression Specification . . . 37
2.4 Results . . . 38
2.5 Conclusions and Limitations . . . 42
2.A Clustering . . . 48
2.B Dependent Variable Breakdown . . . 49
2.C Covariate Descriptive Statistics . . . 50
2.D Original Estimates . . . 52
2.E Software . . . 54

3 `decompr`: Global Value Chain decomposition in `R` **55**
3.1 Introduction . . . 56
3.1.1 Package Details . . . 57

3.2 Data . . . 58
3.3 Leontief decomposition . . . 59
3.3.1 Theoretical derivation . . . 61
3.3.2 Implementation . . . 63
3.3.3 Output . . . 63
3.4 Wang-Wei-Zhu decomposition . . . 64
3.4.1 Theoretical derivation . . . 66
3.4.2 Implementation . . . 68
3.4.3 Output . . . 68
3.5 Conclusion . . . 70

4 Global Value Chains in LICs **74**
4.1 Introduction . . . 74
4.2 New data and new indicators . . . 76
4.2.1 Wang-Wei-Zhu decomposition . . . 77
4.2.2 OECD ICIOs . . . 80
4.3 What we know: Old facts with new data . . . 81
4.4 The role of developing economies: New trends and patterns in GVCs . . . 85
4.4.1 General trends in the GVC participation of developing economies . . . 85
4.4.2 Revealing new trends in the participation of developing economies . . . 87
4.5 Conclusion . . . 92
4.A Output tables . . . 96

Final Remarks **99**

A `rnn`: Recurrent Neural Networks in `R` **100**

B Reproducibe Research Methods **101**

Introduction

As mentioned in the title, methodologically I focus on making the research in this thesis reproducible. Reproducibility is different from replicability in that is refers to regenerating the research results based on the same data, as opposed to replicablity, which uses newly gathered data.

A large part this methodology comes from the field of biostatistics and it is best explained by one of the key figures in this field.

> The replication of scientific findings using independent investigators, methods, data, equipment, and protocols has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study there are examples of scientific investigations that cannot be fully replicated because of a lack of time or resources. In such a situation, there is a need for a minimum standard that can fill the void between full replication and nothing. One candidate for this minimum standard is "reproducible research", which requires that data sets and computer code be made available to others for verifying published results and conducting alternative analyses.

I believe that this applies to at least the same extend an probably more in economics. Although there are situations in which replicable research may be conducted, specifically in experimental settings such as those often used in behavioural economicss or in field research using Randomised Control Trials (RCTs), there are also many cases in which this is not possible.

In this thesis I include two chapters where the identification strategy is based on a natural experiment. Firslty, a change in government policy, as a results of the unconstitutional nature of the sex-based discrimination in pension eligibility in South Africa. Secondly, the introduction of an interface language on the South African Google Search website, as a spillover of that translation work being done for Botswana. In both cases I use the National Income Dynamics Study, the most comprehensive panel data set on South Africa. Since it will be hard to find more relevant data, a full replication - using new data - will prove difficult. As

such, it is all the more important for research to be as transparent as possible, making it at least as reproducible as can be.

There are a number of way in which research can be made reproducible, many of which are already becoming increasingly common. In addition to this, there are several methods which are less widespread, but nevertheless very useful, I discuss these in detail in appendix B.

Chapter 1

Grandfathers and Grandsons: Effects on a Male-Only Pension Change

Abstract

An exogenous male-only increase in cash transfer causes an increase in expenditure on food, an improvement in Weight-for-Age z-scores, a deterioration in Height-for-Age z-scores and no change in Weight-for-Height scores in children in the same household, as observed in the context of South Africa's 2010 state pension expansion for males. When estimated separately, these effects disappear for girls, whereas for boys they remain intact. In 2010 the male eligibility age for the South-African state pension was brought on a par with female eligibility age (60, previously 65). I exploit this policy change in order to estimate the effects of the male-only change in cash transfers, on growth of young children living in the same household, as well as on food expenditure. The policy change took place shortly after the completion of the first wave of South Africa's National Income Dynamics Survey and shortly before the start of the second wave, which lends itself well for a Difference-in-Differences approach on the Right-Hand Side. On the Left-Hand Side I use z-scores of the anthropometrics status of young children in the household (against WHO standards) as well as food expenditure.

1.1 Introduction

Conditional Cash Transfer schemes are increasingly common in developing countries. However, the efficacy of these schemes - in particular in terms of the investments in children - as compared to in-kind benefits such as universal health care are often brought into question, especially when the recipient of the transfer is male.

With this study I seek to answer the question, whether cash transfers can be effective if the recipient is male, specifically, when looking at the effects on the anthropometric status of children in the household, and if these effects are different for boys and girls.

In order to answer this question, I examine a policy change in the South African state pension scheme that only benefits men, namely the bringing down of the eligibility age from 65 years of age to 60 years, on a par with women's eligibility age (Ralston et al., 2015; Skweyiya, 2008).

I find that this policy change leads to increased expenditure on food, improvements in Weight-for-Age, a deterioration in Height-for-Age, and no effect on Weight-for-Height of young children in the household. Furthermore, when estimated separately, for girls the effect disappears, for boys the same effects remain intact at greater significance levels. This suggest that the additional income of newly eligible men only affected boys and not girls. Additionally, it shows that increased expenditure on food has ambiguous effect on anthropometrics.

The debate on Conditional Cash Transfer (CCT) schemes and anthropometric status is closely linked. One of the key determinants of a lag in anthropometric status z-scores is malnutrition. Malnutrition can affect physical and cognitive development, which effects future productivity and income (Dasgupta, 1997; Strauss, 1986). In spite of this, low levels of investment in child health are often observed, despite the inefficiency of this (Arcand, 2001; Rosenzweig and Schultz, 1982, 1983).

Although Conditional Cash Transfer schemes are much more common, it can be more informative to study Unconditional Cash Transfer schemes (UCTs) as these are inherently less prone to selection bias issues. State pension systems are an opportune subject of study, since they are often virtually unconditional upon reaching a certain age.

The South African pension system is of particular interest, because of the relatively high amount of the payout. Upon the initial expansion to include the black population, in 1991, this amount was as much as twice the median monthly income for the rural population (Tangwe and Gutura, 2013). At the time of this study, the state pension payout of is just over 1000 ZAR, which is about half

of median income in South Africa. As a result of this, although the pension system was intended as a form of poverty relief for the elderly population, it has also become that for the South-African general population, serving as a general source of income to many households (Ralston et al., 2015).

Most of the literature suggests that cash transfers to women are more beneficial to children. In particular, Duflo (2000, 2003) studies the effects of South African state pension shortly after the policy change that expended the system to include the black and other non-white populations in 1993. That study finds that income that accrues to a woman in the household leads to improvements in anthropometic status z-scores of girls living in the same household. Studying the grandparents is particularly relevant in the South African context, as they are often the primary caregivers when parents works in other parts of the country. Currently less than one in three children lives with their biological parents (SA Stats, 2016). At the time of this study in 1993, the male eligibility age was 65 and the female age was 60, while average life expectancy in South Africa being significantly lower than that at around 56 (SA Stats, 2016), complicating the comparison. In addition to this, since then, social security coverage has expanded from a rate of 2.5-million in 1994 to over 12.7-million in 2008 (Skweyiya, 2008).

This age discrepancy has always been considered discriminatory on the basis of age and thereby unconstitutional by the South African government. As a consequence of this, in 2010 the pension eligibility age for South African men was lowered from 65 to 60 years old, eliminating the discrepancy (Skweyiya, 2008). I exploit this policy change by estimating a Difference-in-Differences model, quantifying the effects of the male-only increase in cash transfers to households with male household members aged 60 until 65.

This lowering of the pension eligibility age for men, took place between the first and the second wave of data collection for the South African National Income Dynamics Study (Southern Africa Labour and Development Research Unit, 2008, 2012, 2013), which took place respectively in 2008 and 2011. This household survey includes data on age, state-pension eligibility and receipts, children's anthropometric z-scores, income, food/non-food expenditure, etc. The children's z-scores are computed by comparing their anthropometics with the WHO Child Growth Standards (de Onis, 2006; Onyango, 2009; World Health Organization, 2006).

The availability of data directly prior to and after the policy change, enables me to estimate the effect of the cash transfer to the newly eligible group of males aged 60 until 65 in their households using a Difference-in-Differences approach. In particular, I estimate effect of this change on food and non-food expenditure as well as on the anthropometric status z-scores of young children living in the

same households.

I find that the transfer leads to an increase in food expenditure, but shows no significant impact on non-food expenditure. The effects on the anthropometric status of children in the same households are more ambiguous. The change led to an improvement in the Weight-for-Age z-scores, as well as a regression in the Height-for-Age z-scores of children. When I estimate these equations separately for boys and girls, the same variables have no significant effect for girls, for boys the effects remain intact, at greater significance levels.

These results suggest that the increased expenditure in food results in improvements in the short-term Weight-for-Age indicators, but at the same time that the more long-term effect on Height-for-Age is opposite to this. A possible explanation for this is that the increase food expenditure goes towards unhealthy food, increasing weight, but not leading to any long term increases in growth. Furthermore, the separate estimations suggest that this change only affects boys.

The following section §1.2 discusses the National Income Dynamics Study, as well as the WHO Child Growth Standards used to compute the anthropometric z-scores. This is followed by section §1.3 which discusses the empirical model estimated, as well as the tools employed for this. In section §1.4, I present the outcome of these estimations. Finally, I interpret these results and their limitations in section §1.5.

1.2 Data

The Southern Africa Labour and Development Research Unit (NIDS, 2008, 2012, 2013), collected the data in the National Income Dynamics Study, together with the World Bank, which is the primary source of data in this study.

The study collects information on a representative set of approximately ten thousand South-African households over time. Currently three 'waves' of data are available, these waves date from 2008, 2011, and 2013. The key variables that I use are that I use are:

- child anthropometrics status, Weight for Age, Weight for Height, and Height for Age (`zwfa`, `zwfh`, `zhfa`);
- food and non-food expenditure (`expf`, `expnf`);
- child age in days (`c_age_days1`);
- gender of the child (`woman`);
- pension eligible adult (`man_60_65`, `woman_60_65`, `man_65`, `woman_65`).

In addition to these variables of interest, I include a number relevant of covariates, such as household income (`hhincome`), in the analysis.

Children's anthropometrics including length/height, weight, and waist, are combined with the WHO growth standards, to calculate the z-scores. In table 1.2 and table 1.1 the distributions of both variables for both boys and girls are presented. table 1.1 presents distributions statistics on the food expenditure variable.

Figure 1.1: Weight-for-Age z-score distributions

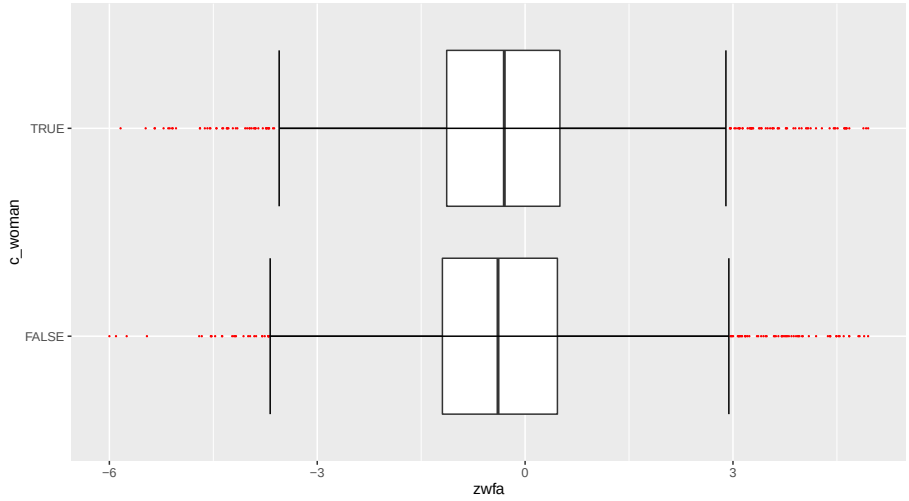
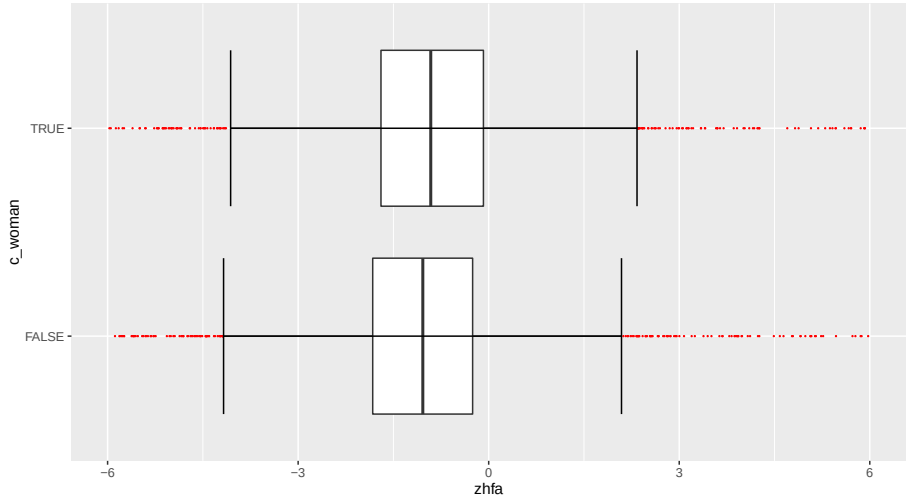


Figure 1.2: Height-for-Age z-score distributions



For adults several variables measure the different amounts and sources of income. Among those, a variable if the adult receives a state pension, and if

so, how much. This is a numeric variable, the values of which lie very close together. A total of 22.99% of households in the dataset have a female pension recipient as a household member. Conversely, only 9.06% of households in the dataset have a male pension recipient as a household members.

The income received by state pension recipients is generally the same amount, which slightly more than 1000 ZAR. This is about half of median household income, as recorded in this dataset.

I construct a dummy variable for children living in a household with a man aged 60 until 65 (`man_60_65`). As well as dummies for a man 65 years or older (`man_65`) and these same dummies for children living with women of those ages. The coefficient of interest relates to the interaction of the `man_60_65` dummy variable with the `event` dummy, whereby the interaction with the additional household member dummies serve to distinguish the outcome.

In addition to the explanandum, I include number of relevant covariates on the Right-Hand Side (RHS). table 1.2 gives a description of the distribution of household income as found in the dataset.

In 2006 the WHO published its standards for child growth(de Onis, 2006), superseding the previously used CDC Growth Charts of (Kuczmarski, 2000, CDC Growth Charts: United States). The WHO charts map the average growth of an ethnically varied sample of children living in households with healthy lifestyles, setting a benchmark for growth.

The various anthropometric z-scores are representative of child health in related but somewhat different ways. The overarching idea is that factors that harm the child's health, such as malnutrition on illness, adversely affects these variables. Both malnutrition and illness can cause the child to grow less for a certain amount of time, as a result, the height of the child is on average less than that other children of the same age that have not suffered this period of malnutrition or disease. This effect not disappear over time, making Height-for-

Table 1.1: Food expenditure

wave	minimum	q1	median	mean	q3	maximum	NA's
1	24	500	730	947	1148	14780	NA
2	33	560	841	1015	1219	27380	1456
3	30	600	820	1061	1216	30000	944

Table 1.2: Income distribution

wave	minimum	q1	median	mean	q3	maximum	NA's
1	0.0	1284	2165	4014	3966	130000	NA
2	100.0	1500	2583	4720	4817	446900	1089
3	126.2	1980	3376	5541	5933	300200	944

Age an indicator that can reflect past malnutrition or illness, even after quite some time. The Weight-for-Height is an indicator that relates children's weight to their height, irrespective of their age. Weight can also be adversely affected by malnutrition or disease, but unlike Height-for-Age, this effect disappears of time in healthy periods.

For instance, if we measure a height x for a child of age y (in weeks/months), then we refer the to WHO tables, find the relevant ideal height and standard deviation for a child of age y. We then subtract the ideal height (μ_y) from the observed height, and divide by the standard deviation (σ_y), like so:

$$z_{xy} = \frac{x - \mu_y}{\sigma_y}$$

These ideal scores are based on a sample of children from different ethnic populations, in households which observed a healthy lifestyle. Any health issues, such as malnutrition or disease will affect these metrics, by causing the child to be shorter or lighter as compared to these ideal standards. It is however not possible to distinguish between the different causes of an observed slowed growth.

It is best practice to use only metrics for children between the ages of 6 months and 60 months.

Here we use two type of z-scores, the Height-for-Age Z-score (HAZ) and the Weight-for-Age Z-score (WAZ). Since these metrics are both age-based, they provide information about all past growth issues. Any past issues such a malnutrition and disease will have impaired growth, and these effects will still be captured by today's height. This also applies to the WAZ, as the ideal weight is a function of the height, which is in turn a function of the age.

These are constructed on a weekly basis up to the age of 60 months, and on a monthly basis thereafter.

The NIDS uses a file and data structure which is ill suited for panel data analysis. I therefore transform the data to a format which is more conducive to panel data analysis. For this I used 'Tidy Data' structure, as described in Wickham (2014).

1.3 Empirical methodology

With this study I aim to answer the question whether a male-only cash can be effective, in particular with regard to the effects on the growth of young children in the same household.

1.3.1 Identification strategy

The identification strategy in this paper is based on a policy change in the pension eligibility age for men. Until the middle of 2009, men became eligible for the state pension at the age of 65. Between mid 2009 and December 31st 2010, this was gradually lowered to 60 (Ralston et al., 2015). I combine this information with data from the South-African National Income Dynamics Study, a full-panel dataset, which contains information on households from before and after this policy change. This policy change falls between waves 1 and 2 of the NIDS, which took place in 2008 and 2011 respectively.

I study the effect of this policy change, on the anthropometric status of children in the same household as pension recipients as well as on food and non-food expenditure.

This identification strategy is operationalised by constructing a policy change or event dummy. This event dummy is called `event`, and takes the value `1` for data after the policy change (waves 2 & 3), and the value `0` before the policy change (wave 1). This dummy is interacted with a dummy variable on having a male household member aged between 60 and 65 (`man_60_65`), as well as with dummies for men age 65 or older and dummies for women of the same ages.

1.3.2 Regression specification

In order to identify the effect of the policy change, I employ Difference-in-Differences, using the fixed-effects estimator (the Hausman test rejected random effects as biased, see section §1.A). Based on this setup, I formulate two base models. One model with the `event` dummy, and an interaction term with male pension recipient (`man_60_65`). The second model has the `event` dummy and an interaction term with the same `man_60_65` dummy, as well as interaction terms with the dummies for women between 60 and 65 and women and men 65 and above. Each of these models is estimated with both types of z-scores, as well as food and non-food expenditure as dependent variables.

The outcome variable is y_{it}, this outcome variable takes the form of the of the z-scores, such as Weight-for-Age or Height-for-Age, food/non-food expenditure. Here t denotes time and i the individual. The individual and time fixes effects are denoted by γ_i and λ_t respectively. Dummies for living in a household with a female or a male pension recipient are included as P_{it}^f and P_{it}^m respectively. The dummy variable T_{it} denoted the treatment status. Lastly, ϵ_{it} is the error term, which is assumed to be distributed as $\epsilon_{it} \sim N(0, \sigma)$.

$$y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \beta X_{it} + \epsilon_{it} \tag{1.1}$$

In this, α_i represents the individual fixed effects, λ_t represent the time fixed effects, and X_{it} are the time varying covariates. The error term is ϵ_{it}. Finally the term of interest is D_{it} , which is the interaction of the `man_60_65` dummy variable with the `event` dummy variable. Whereby the coefficient of interest being δ. In the second specification, I also include dummy variables for male household members of 65 years and over, and for woman aged 60-65 and 65 years and over. In this specification, D_{it} represent the interaction of the vector of household member dummies, with the event dummy.

The event variable `event` is a dummy which takes the value `TRUE` (i.e. `1`) for the data collected after the policy change, i.e. waves 2 and 3 and `FALSE` (i.e. `0`) for data collected before then. Lastly, I include the covariate `hhincome` which represents total household income.

As the variables of interest living with a man between 60 and 65, as well as the other household member dummy variables are determined at a household level, I apply standard error corrections to take this into account, the full procedure is outlined in section §1.A.

As described above, I use a total of five dependent variables, Height-for-Age (HAZ), Weight-for-Height (HAZ), Weight-for-Age (WAZ), and food and non-food expenditure. Each of these is used in a different estimation as the Left-Hand Side (LHS) variable. Combining these five LHSs with each of the RHS specifications. I present the key results of these estimations in section §1.4.

1.4 Results

In table 1.3 I present the result of the estimations with food and non-food expenditure as dependent variables. In table 1.4 and table 1.5 I present the estimation results for the equations with anthropometric status as the dependent variables, coefficients with a corresponding p-value of less than 0.1 are printed in bold font.

The first four items in each table represent the interaction terms between the event dummy and the various household member dummies, e.g. the variable `man_60_65` is a dummy variable for the child living in the same household as a male aged 60 until 65.

The other rows represent the independent variables. Where `woman_60_65` represents the dummy variable for children living in a household with a state pension eligible woman aged 60 until 65. The variables `man_65` and `woman_65` represent pension eligible men an women over the age of 65 respectively.

As mentioned in section §1.3 I estimate the equations using only the interaction term with `man_60_65` as well as with all household member of age dummies, such as woman_65, etc. I only present the results here using all in-

teraction terms. The estimates with the sole interaction term are qualitatively and quantitatively similar to the full estimates and are available in appendix 1.A.

The key result in table 1.3 is that in the estimation with food expenditure as the explanandum the coefficient estimate for the interaction term event * man_60_65, is positive at 103.82 with a corresponding p-value of 0.03. This coefficient estimate represents a 103.82 Rand (ZAR) increase on food expenditure in the households which received the additional income through pension receipts of men aged between 60 and 65. The interaction term with Woman 65+ is also positive and significant, although as we will see below, this effect does not permeate to any of the anthropometric z-scores. In addition to this, the estimate for household income is positive at 0.03 and highly significant at 0.00.

In the estimation with non-food expenditure as the explanandum, I find so significant effects other than household income.

Table 1.3: Food and Non-Food Expenditure

	Food	(P > |t|)	Non-Food	(P > |t|)
event * Man 60-65	**103.82**	(0.03)	-0.05	(0.22)
event * Man 65+	0.09	(1.00)	-0.02	(0.35)
event * Woman 60-65	4.69	(0.90)	0.02	(0.36)
event * Woman 65+	**104.24**	(0.00)	0.01	(0.51)
Man 60-65	55.61	(0.20)	206.16	(0.57)
Man 65+	**112.18**	(0.00)	377.43	(0.57)
Woman 60-65	46.88	(0.13)	-356.02	(0.16)
Woman 65+	**-0.68**	(0.00)	-181.57	(0.38)
event	35.23	(0.00)	131.32	(0.08)
Household Income	**0.03**	(0.00)	**0.09**	(0.00)
Observations	15938		15938	

When using Weight-for-Age as a dependent variable, the coefficient of interest is positive at 0.33 and significant, with a p-value of 0.02. This represents an increase of 33% of a standard deviation of the WHO Child Growth Standards towards the WHO target mean value.

The interaction terms of the event dummy with the other household member dummies, `man_65`, `woman_60_65`, and `woman_65`, are not significant with p-values of 0.74, 0.18, and 0.58 respectively.

Household income is positive and highly significant, although effect of one additional Rand on the z-scores is very small. The coefficient of Girl is position at 0.09 and significant, indicating that girl's have a lag 9% less of a WHO Standards standard deviation.

I also estimate this equation separately for boys and for girls. The estimation

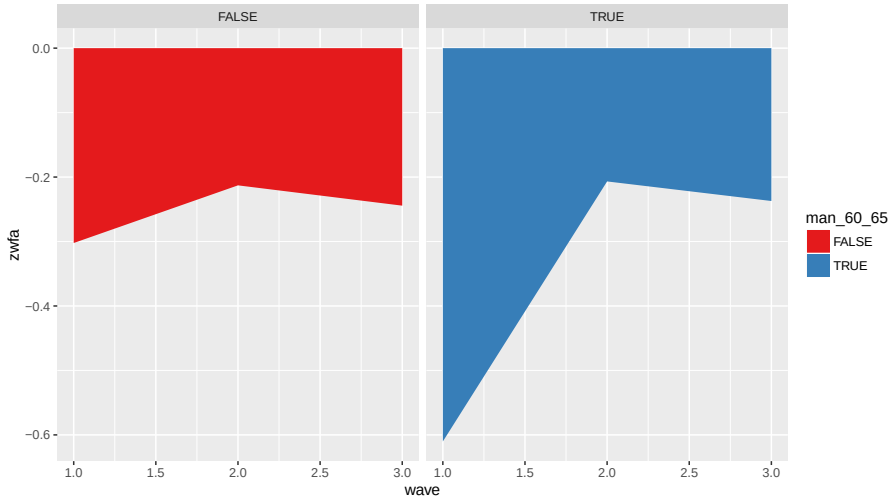
including only boys gives a somewhat higher estimate of 0.44, with a p-value of 0.03. The estimate for girls is lower at 0.26 and insignificant at a p-value of 0.23.

A graphical illustration of the result on the variable of interest can be found in figure 1.3, which shows the average lag in `zwfa` for children living with a man between 60 and 65, before and after the policy change on the right, as compared to children who do not have such a household member on the left.

Table 1.4: Weight for Age

	General	(P > |t|)	Boys	(P > |t|)	Girls	(P > |t|)
event * Man 60-65	**0.33**	(0.02)	**0.44**	(0.03)	0.26	(0.23)
event * Man 65+	-0.03	(0.74)	0.15	(0.31)	**-0.30**	(0.05)
event * Woman 60-65	0.15	(0.18)	0.18	(0.23)	0.13	(0.44)
event * Woman 65+	0.04	(0.58)	0.29	(0.79)	0.03	(0.77)
Man 60-65	**-0.27**	(0.03)	**-0.35**	(0.06)	-0.29	(0.13)
Man 65+	-0.06	(0.47)	**-0.35**	(0.01)	**0.26**	(0.04)
Woman 60-65	**-0.17**	(0.06)	-0.14	(0.27)	-0.18	(0.21)
Woman 65+	0.03	(0.69)	0.09	(0.32)	-0.02	(0.80)
event	**0.00**	(0.00)	0.01	(0.85)	0.06	(0.19)
Household Income	**0.00**	(0.00)	**0.00**	(0.00)	0.00	(0.00)
Girl	**0.09**	(0.00)				
Observations	11740		5878		5862	

Figure 1.3: Weight for Age



When I use Height-for-Age as a dependent variable, the result is very different. I find a negative effect of -0.52 of the treatment on the dependent variable, at a p-value of 0.09. This represents a drop of 52% of a standard deviation in the WHO Child Growth Standards, falling further below the WHO target mean

value.

None of the interaction terms with any of the other household member dummies are significant, at p-values of 0.31, 0.24, and 0.98 respectively. When I re-estimate this equation to only include boys in the sample, I find a much greater effect of -0.94, which is also more significant, with a p-value of 0.04. The same is not true of the re-estimation which only includes girls, where the estimate is -0.24 and entirely insignificant with a p-value of 0.57.

The coefficient estimate for Girl is 0.22, which means that on average across all three waves, girls z-scores are 22% of a WHO standard deviation less below the WHO target mean value than boys are. Again, the value of the household income coefficient is very low (rounded to 0.00) but highly significant at a p-value of 0.00, this is simply because the marginal effect of one additional Rand is very small but nevertheless decisively positive.

Table 1.5: Height for Age

	General	(P > |t|)	Boys	(P > |t|)	Girls	(P > |t|)
event * Man 60-65	**-0.52**	(0.09)	**-0.94**	(0.04)	-0.24	(0.57)
event * Man 65+	-0.23	(0.31)	-0.03	(0.92)	-0.36	(0.25)
event * Woman 60-65	0.27	(0.24)	0.55	(0.08)	0.02	(0.95)
event * Woman 65+	-0.00	(0.98)	-0.16	(0.49)	0.14	(0.54)
Man 60-65	0.11	(0.66)	0.35	0.37	-0.06	(0.85)
Man 65+	0.19	(0.29)	-0.07	(0.79)	0.42	(0.08)
Woman 60-65	**-0.32**	(0.08)	-0.26	(0.32)	-0.40	(0.15)
Woman 65+	0.04	(0.74)	0.12	(0.50)	-0.03	(0.88)
event	-0.01	(0.867)	0.01	(0.92)	-0.03	(0.77)
Household Income	**0.00**	(0.00)	0.00	(0.00)	**0.00**	(0.00)
Girl	**0.22**	(0.00)				
Observations	4809		2301		2377	

1.5 Conclusions and limitations

The impetus for this paper is to gain a greater understanding of the efficiency of cash transfers to male household members, to for instance improve the designs of CCTs. To do this I seek to answer the question if cash transfers to men are effective in influencing the health of household members, particularly the growth of children in the same household.

I analyse data from the National Income Dynamics Study around the exogenous lowering of the pension eligibility age for men, from 65 to 60, in the South African state pension system.

Using a Difference-in-Differences based estimation, I compare differences in the anthropometric status of children living in the same household as men of the newly eligible age with children live in a household without such a member.

On the Right-Hand Side I use a dummy variable for the policy change as well as dummy variables for men aged 60-65, men aged 65+ and the same two dummy variables for women of those ages as household members. On the Left-Hand Side (LHS) I use three different anthropometric z-scores, Height-for-Age, Weight-for-Age, and Weight-for-Height, as well as food and non-food expenditure. In addition to the general estimates using z-scores as a dependent variable, I also re-estimate these three equations once using only boys in the sample and once using only girls.

The estimation of these equations gives three key results. Firstly, I find that there is a significant and consistent positive effect of the interaction term of the event dummy and the `man_60_65` dummy, on food expenditure. Secondly, there is a significant and consistent positive effect of the same interaction term on the Weight-for-Age Z-score, the interaction terms with other household member dummies are not significant. Thirdly, I find a consistent and negative effect of the interaction term on the Height-for-Age Z-scores. I find no significant effect on Weight-for-Height. All of these effects are consistent across the different specifications and use standard error corrections.

These effects suggest that the male income resulted in greater food expenditure, which improved the Weight-for-Height anthropometric status. Since the estimation includes data from two waves after the policy change, we would expect this additional food expenditure to also lead to an improvement in the more long-term metric of Height-for-Age. Surprisingly we observe an opposite effect here. An explanation for this could be that nature of the expenditure on food has changed, helping improve the Weight-for-Age, but not improving the Height-for-Age, with the later being a more long term health indicator.

When I subset to a data to only includes girls and reestimate the equation with the two z-score dependent variables, I find no significant effects. However, using a subset of only boys gives the same results as in the original estimation and at greater significance levels. This makes it unlikely that the surprising deterioration in Height-for-Age can be attributed to another factor such as the 2008 global financial crisis, since this would in all likelihood also have affected girls as it did boys.

More central to this study, it suggest that the cash transfer only affected boys living with a new recipient. As shown in table 1.1, the boys in the dataset scored slightly worse in the Weight-for-Age z-scores before the policy change. One possible explanation for the improvement in Weight-for-Age in boys could be that they benefitted more in terms of this metric because the addition food expenditure had a relatively large effect due as a result of the initial greater lag. However, this is at odds with the deterioration in Height-for-Age, since also here, boys scored worse even in wave 1 of the dataset. Should a greater lag

have been the cause of the effect in the improvement in Weight-for-Age then we would expect a similar improvement in Height-for-Age.

An alternative explanation could be that if the cash transfer was used to purchase unhealthy food, which increases weight but is not nutritious in terms of promoting growth, than this was only spent on boys in the household. If there is merit to this last explanation, then this result mirrors the result found in Duflo (2000, 2003), where the central result is an improvement in girls' anthropometric status when they live with a female pension recipient. The combination of these two results is in that sense similar to Thomas (1994), who finds that within households, father's education has a larger effect on the anthropometric status of boys and mother's education has a larger effect on the anthropometric status of girls.

The first key outcome of this research is that food expenditure is some way too ambiguous a variable. The results of the increased food expenditure are not altogether positive, leading to a deterioration in Height-for-Age. In context such as these, the increased food expenditure would generally be thought of as a positive development. By further analysing the types of food purchased, it might be possible to better understand the deterioration in Height-for-Age and the lack of an effect on Weight-for-Height.

The second key outcome of this research is that both the positive and negative consequence of the male-only cash transfer seem to only effect boys. This could suggest that, similarly to Duflo (2000, 2003) the grandparent seems to spend only on grandchildren of the same sex.

Bibliography

Arcand, Jean-Louis

2001 *Undernourishment and economic growth: the efficiency cost of hunger*, 147, Food & Agriculture Org.

Croissant, Yves

2013 "pglm: Panel Generalized Linear Model", *R package version 0.1-2*, http://CRAN.%20R-project.%20org/package=%20pglm.

Croissant, Yves, Giovanni Millo et al.

2008 "Panel data econometrics in R: The plm package", *Journal of Statistical Software*, 27, 2, pp. 1–43.

Dasgupta, Partha

1997 "Nutritional status, the capacity for work, and poverty traps", *Journal of Econometrics*, 77, 1, pp. 5–37.

De Onis, Mercedes et al.

2006 *WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development*, World Health Organization, http://www.who.int/childgrowth.

Duflo, Esther

2000 "Child health and household resources in South Africa: Evidence from the Old Age Pension program", *The American Economic Review*, 90, 2, pp. 393–398, http://www.jstor.org/discover/10.2307/117257.

2003 "Grandmothers and Granddaughters: Old-Age Pensions and Intrahousehold Allocation in South Africa", *The World Bank Economic Review*, 17, 1, pp. 1–25, DOI: 10.1093/wber/lhg013.

Git Team

2016 *Git: Software Code Manager*, 137 Montague ST STE 380, Brooklyn, NY 11201-3548, `http://www.git-scm.org/`.

Kuczmarski, RJ et al.

2000 *CDC growth Charts: United States*, 314, pp. 1–28, `http://www.cdc.gov/growthcharts/reports.htm`.

Onyango, AW

2009 "[World Health Organization child growth standards: background, methodology and main results of the Multicentre Growth Reference Study]." *Archives de pediatrie: organe officiel de la Societe francaise de pediatrie*, 16, 6, pp. 735–736.

R Core Team

2016 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`.

Ralston, Margaret, Enid Schatz, Jane Menken, Francesco Xavier Gomez-Olive and Stephen Tollman

2015 "Who Benefits - Or Does not - From South Africa's Old Age Pension? Evidence from Characteristics of Rural Pensioners and Non-Pensioners", *International journal of environmental research and public health*, 13, 1, p. 85.

Rosenzweig, Mark R and T Paul Schultz

1982 "Market opportunities, genetic endowments, and intrafamily resource distribution: Child survival in rural India", *The American Economic Review*, 72, 4, pp. 803–815.

1983 "Estimating a household production function: Heterogeneity, the demand for health inputs, and their effects on birth weight", *The Journal of Political Economy*, pp. 723–746.

SA Stats

2016 "Statistics South Africa", *Stats in Brief*.

Skweyiya, Zola

2008 *SA men get state pensions earlier*, `http://www.southafrica.info/services/government/pension-160708.htm` (visited on 13/06/2016).

Southern Africa Labour and Development Research Unit

2008 *National Income Dynamics Study, Wave 1*, version 5.1, `http://www.nids.uct.ac.za/home/`.

2012 *National Income Dynamics Study, Wave 2*, version 2.1, `http://www.nids.uct.ac.za/home/`.

2013 *National Income Dynamics Study, Wave 3*, version 1.1, `http://www.nids.uct.ac.za/home/`.

Strauss, John

1986 "Does better nutrition raise farm productivity?", *The Journal of Political Economy*, pp. 297–320.

Tangwe, Pius Tanga and Priscilla Gutura

2013 "The Impact of the Old Age Grant on Rural Households in Nkonkobe Municipality in the Eastern Cape Province of South Africa", *Mediterranean Journal of Social Sciences*, 4, 13, p. 627.

Thomas, Duncan

1994 "Like father, like son; like mother, like daughter: Parental resources and child height", *Journal of Human Resources*, pp. 950–988.

Wickham, Hadley

2014 "Tidy Data", *Journal of Statistical Software*, 59, 1, pp. 1–23, DOI: `10.18637/jss.v059.i10`, `http://www.jstatsoft.org/index.php/jss/article/view/v059i10`.

2016 *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, R package version 0.4.1, `https://CRAN.R-project.org/package=tidyr`.

Wickham, Hadley and Romain Francois

2015 *dplyr: A Grammar of Data Manipulation*, R package version 0.4.3, `https://CRAN.R-project.org/package=dplyr`.

World Health Organization et al.

2006 *World Health Organization child growth standards*.

1.A Original Estimates

Table 1.6: Standard Error Correction

```
library(lmtest)   # standard error correction
library(broom)    # output formatting

# standard error correction
tidy( coeftest(expf1, vcov=vcovHC(expf1,
                                  type="HC0",
                                  cluster="group")) )
```

term	estimate	std.error	statistic	p.value
(Intercept)	797.973137	20.4834004	38.957064	0.0000000
post_treatmentTRUE	54.598225	9.4195865	5.796244	0.0000000
man_60_65TRUE	60.992207	30.8335999	1.978109	0.0479254
man_65TRUE	112.135045	19.9611804	5.617656	0.0000000
woman_60_65TRUE	47.634524	15.3353771	3.106185	0.0018969
woman_65TRUE	3.158883	12.1322850	0.260370	0.7945802
hhincome	0.034387	0.0047316	7.267585	0.0000000
womanTRUE	-12.610512	10.1037207	-1.248106	0.2120019
post_treatmentTRUE:man_60_65TRUE	97.758973	39.4529129	2.477864	0.0132225

Table 1.7: Non-food expenditure

```
NIDS %>%
  group_by(wave) %>%
  do(tidy(summary(.$expnf)))
```

wave	minimum	q1	median	mean	q3	maximum	NA's
1	4.000	220.0	552.4	1789	1425	120300	NA
2	1.000	285.1	588.1	1678	1300	361000	1456
3	4.429	336.0	755.0	1870	1735	112000	944

Table 1.8: Non-Food Expenditure

```
plm(expnf ~     post_treatment*post_treatment +
                man_65 +
                wopost_treatment +
                woman_65 +
                hhincome +
                woman,
                data   = NIDS,
                model  = 'random')
```

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	1052.4112911	56.710077	18.5577476	0.0000000
post_treatmentTRUE	-229.1189649	53.971673	-4.2451707	0.0000219
man_60_65TRUE	-12.9406582	245.515119	-0.0527082	0.9579648
man_65TRUE	-345.7800113	96.666790	-3.5770300	0.0003481
woman_60_65TRUE	-385.0294955	93.983693	-4.0967692	0.0000420
woman_65TRUE	-617.4061057	70.636448	-8.7406166	0.0000000
hhincome	0.2310988	0.003067	75.3510870	0.0000000
womanTRUE	-52.8786218	54.028761	-0.9787125	0.3277298
post_treatmentTRUE:man_60_65TRUE	-231.9316003	281.149594	-0.8249402	0.4094120

Table 1.9: Food Expenditure Interact All

```
plm(expf ~      post_treatment*post_treatment +
                post_treatment*man_65 +
                post_treatment*wopost_treatment +
                post_treatment*woman_65 +
                hhincome +
                woman,
                data   = NIDS,
                model  = 'within')
```

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	810.8105328	10.8755576	74.5534680	0.0000000
post_treatmentTRUE	35.2251442	10.5897866	3.3263318	0.0008810
man_60_65TRUE	55.6196875	43.0213993	1.2928377	0.1960770
man_65TRUE	112.1835241	28.5781790	3.9254959	0.0000867
woman_60_65TRUE	46.8846693	30.9664240	1.5140485	0.1300239
woman_65TRUE	-67.8984723	21.0318516	-3.2283640	0.0012463
hhincome	0.0344025	0.0005493	62.6291630	0.0000000
womanTRUE	-12.4862470	10.2886736	-1.2135915	0.2249131
post_treatmentTRUE:man_60_65TRUE	103.8232181	49.1094384	2.1141194	0.0345132
post_treatmentTRUE:man_65TRUE	0.0905759	33.1925710	0.0027288	0.9978228
post_treatmentTRUE:woman_60_65TRUE	4.6874243	35.5632106	0.1318054	0.8951391
post_treatmentTRUE:woman_65TRUE	104.2408234	24.1584808	4.3148749	0.0000160

Table 1.10: Girls Height for Age

```
plm(zhfa ~      post_treatment*man_60_65 +
                post_treatment*man_65 +
                post_treatment*woman_60_65 +
                post_treatment*woman_65 +
                hhincome,
                data = NIDS,
                subset = best_age_yrs < 4 &
                woman == TRUE,
                model='between')
```

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-1.0827829	0.0853476	-12.6867426	0.0000000
eventTRUE	-0.0292989	0.1014385	-0.2888346	0.7727352
man_60_65TRUE	-0.0656365	0.3501531	-0.1874508	0.8513245
man_65TRUE	0.4217243	0.2468947	1.7081136	0.0877565
woman_60_65TRUE	-0.3987542	0.2798468	-1.4249014	0.1543277
woman_65TRUE	-0.0273289	0.1885614	-0.1449338	0.8847765
hhincome	0.0000250	0.0000058	4.3324668	0.0000154
eventTRUE:man_60_65TRUE	-0.2353016	0.4120205	-0.5710920	0.5679957
eventTRUE:man_65TRUE	-0.3603559	0.3146522	-1.1452513	0.2522298
eventTRUE:woman_60_65TRUE	0.0214524	0.3339310	0.0642420	0.9487834
eventTRUE:woman_65TRUE	0.1394328	0.2329406	0.5985769	0.5495168

Table 1.11: Boys Height for Age

```
plm(zhfa ~      post_treatment*man_60_65 +
                post_treatment*man_65 +
                post_treatment*woman_60_65 +
                post_treatment*woman_65 +
                hhincome,
                data = NIDS,
                subset = best_age_yrs < 4 &
                woman == FALSE,
                model='between')
```

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-1.2998129	0.0848152	-15.3252272	0.0000000
eventTRUE	0.0103725	0.1028205	0.1008795	0.9196557
man_60_65TRUE	0.3491800	0.3913368	0.8922751	0.3723477
man_65TRUE	-0.0665181	0.2569582	-0.2588674	0.7957629
woman_60_65TRUE	-0.2599733	0.2609596	-0.9962203	0.3192579
woman_65TRUE	0.1223980	0.1829573	0.6689976	0.5035705
hhincome	0.0000156	0.0000065	2.3876923	0.0170425
eventTRUE:man_60_65TRUE	-0.9431906	0.4647851	-2.0293047	0.0425532
eventTRUE:man_65TRUE	-0.0319730	0.3231811	-0.0989320	0.9212017
eventTRUE:woman_60_65TRUE	0.5501036	0.3207639	1.7149798	0.0864965
eventTRUE:woman_65TRUE	-0.1619634	0.2326134	-0.6962771	0.4863324

Table 1.12: Height for Age

```
plm(zhfa ~      post_treatment*man_60_65 +
                post_treatment*man_65 +
                post_treatment*woman_60_65 +
                post_treatment*woman_65 +
                hhincome +
                woman,
                data = NIDS,
                subset = best_age_yrs < 4,
                model='between')
```

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-1.3045248	0.0661855	-19.7101308	0.0000000
post_treatmentTRUE	-0.0118690	0.0722069	-0.1643745	0.8694440
man_60_65TRUE	0.1125166	0.2609687	0.4311499	0.6663809
man_65TRUE	0.1873098	0.1780119	1.0522319	0.2927522
woman_60_65TRUE	-0.3244063	0.1907804	-1.7004174	0.0891246
woman_65TRUE	0.0430632	0.1312351	0.3281375	0.7428237
hhincome	0.0000211	0.0000043	4.8737618	0.0000011
womanTRUE	0.2245553	0.0563309	3.9863642	0.0000682
post_treatmentTRUE:man_60_65TRUE	-0.5174950	0.3081696	-1.6792542	0.0931751
post_treatmentTRUE:man_65TRUE	-0.2266029	0.2252792	-1.0058756	0.3145319
post_treatmentTRUE:woman_60_65TRUE	0.2697230	0.2306746	1.1692789	0.2423560
post_treatmentTRUE:woman_65TRUE	-0.0021065	0.1644171	-0.0128120	0.9897784

Table 1.13: Weight for Age

```
plm(zwfa ~      post_treatment*man_60_65 +
                post_treatment*man_65 +
                post_treatment*woman_60_65 +
                post_treatment*woman_65 +
                hhincome +
                woman,
                data = NIDS,
                subset = c_age_days1 > 180 & c_age_days1 < 2920,
                model='between')
```

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-0.3073009	0.0326055	-9.4248055	0.0000000
post_treatmentTRUE	0.0032539	0.0326791	0.0995698	0.9206876
man_60_65TRUE	-0.2768796	0.1304236	-2.1229253	0.0337810
man_65TRUE	-0.0623759	0.0854461	-0.7300029	0.4654030
woman_60_65TRUE	-0.1747761	0.0940851	-1.8576383	0.0632454
woman_65TRUE	0.0254201	0.0639132	0.3977282	0.6908378
hhincome	0.0000100	0.0000016	6.2361148	0.0000000
womanTRUE	0.0953099	0.0293109	3.2516874	0.0011505
post_treatmentTRUE:man_60_65TRUE	0.3300113	0.1479880	2.2299866	0.0257672
post_treatmentTRUE:man_65TRUE	-0.0333463	0.1008261	-0.3307311	0.7408535
post_treatmentTRUE:woman_60_65TRUE	0.1452480	0.1078110	1.3472466	0.1779269
post_treatmentTRUE:woman_65TRUE	0.0412345	0.0738919	0.5580384	0.5768288

Table 1.14: Food Expenditure

```
plm(expf ~      post_treatment*man_60_65 +
                post_treatment*man_65 +
                post_treatment*woman_60_65 +
                post_treatment*woman_65 +
                hhincome +
                woman,
                data   = NIDS,
                model  = 'within')
```

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	810.8105328	10.8755576	74.5534680	0.0000000
post_treatmentTRUE	35.2251442	10.5897866	3.3263318	0.0008810
man_60_65TRUE	55.6196875	43.0213993	1.2928377	0.1960770
man_65TRUE	112.1835241	28.5781790	3.9254959	0.0000867
woman_60_65TRUE	46.8846693	30.9664240	1.5140485	0.1300239
woman_65TRUE	-67.8984723	21.0318516	-3.2283640	0.0012463
hhincome	0.0344025	0.0005493	62.6291630	0.0000000
womanTRUE	-12.4862470	10.2886736	-1.2135915	0.2249131
post_treatmentTRUE:man_60_65TRUE	103.8232181	49.1094384	2.1141194	0.0345132
post_treatmentTRUE:man_65TRUE	0.0905759	33.1925710	0.0027288	0.9978228
post_treatmentTRUE:woman_60_65TRUE	4.6874243	35.5632106	0.1318054	0.8951391
post_treatmentTRUE:woman_65TRUE	104.2408234	24.1584808	4.3148749	0.0000160

Table 1.15: Hausmann

```
model_expf2 <- expnf ~ post_treatment*man_60_65 +
post_treatment*man_65 +
woman_60_65*post_treatment +
post_treatment*woman_65 +
hhincome +
woman

fe_expf2 <- plm(model_expf2, data=NIDS, model='within')

## series fwag, cwag, swag, chld, fost, spen_flg, ppen_flg, uif, remt are NA and have been
## series spen, ppen are constants and have been removed

re_expf2 <- plm(model_expf2, data=NIDS, model='random')

## series fwag, cwag, swag, chld, fost, spen_flg, ppen_flg, uif, remt are NA and have been
## series spen, ppen are constants and have been removed

phtest(fe_expf2, re_expf2)

##
##  Hausman Test
##
## data:  model_expf2
## chisq = 857.92, df = 10, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

1.B Software

The computation estimation of these models is performed using R (R Core Team, 2016), with the implementation of the panel data structure and models using the `plm` package by Croissant and Millo (2008). Generalized Linear Models for the panel data set are estimated using the `pglm` package by Croissant (2013).

All changes are logged using the version control system Git (Git Team, 2016) and publicly available on GitHub at https://github.com/bquast/Grandfathers-Grandsonds/[1].

In order to merge data and compute of these statistics, I make use of the dplyr and tidyr R packages Wickham (2016); Wickham and Francois (2015). After having combined the various data.frames within each wave, the three waves can be combined by simply joining the rows using base R's `rbind()` function(R Core Team, 2016).

[1]The repository can by cloned to a local computer by entering in following command in a terminal (with Git installed):
`git clone https://github.com/bquast/Grandfathers-Grandsons.git`

Chapter 2

Making the Next Billion Demand Access

The Local-Content Effect of `google.co.za` in Setswana

Abstract

Recent attempts to connect the current 'next billion' to the Internet in places such as sub-Saharan Africa have not met expectations. In places where Internet infrastructure has come online and prices have gone down, the expected consequent increase in uptake was not observed. Internet adoption in a certain language is a two-sided market with positive cross-side network effects. As a result of this, it is difficult to isolate the causal effect of one on the other. The exogenous introduction of the Setswana language interface on the South African Google Search website was a spillover of the development of that interface for the Botswanan Google website. This exogenous improvement in the accessibility of Setswana-language content has resulted in a substantial increase in the number of native Setswana speakers coming online and owning personal computers. This is turn has also led to increased usage of the Setswana language online. This adoption appears to also lead to improvements in employment.

2.1 Introduction

Internet uptake is a two-sided market, with users on one side and content creators on the other side. Positive cross-side network effects mean that increases in content leads to increases in user adoption and visa versa. This market exists separately for each language but for many indigenous languages this virtuous

circle fails to start properly, keeping usage and content levels low.

With this study I seek to answer the question whether an increase in local language content, does indeed lead to an increase in uptake of Internet usage among native speakers of this language.

Because of the cross-side network effects in a two-sided market, any observed changes are inherently endogenous. I remedy this problem by using an exogenous shock in accessibility of Setswana language content in South Africa, namely the introduction of the Google Search interface in Setswana. I find that this leads to a strong increase in both the proportion of households reporting to have spent on Internet access in the last 30 days, as well as individuals owning a computer. This in turn has led to a large increase in the usage of the Setswana language online (in Google search queries). There also appears to be a strong improvement in employment status among individuals who spend on Internet or own a computer after the introduction of the interface. Suggesting that the expanded demographic of Setswana Internet users is benefiting from increased Internet adoption in terms of employability.

The term 'Connecting the Next Billion' was introduced in The Economist's 2006 'End of Year Report' (Standage, 2006), discussing the infrastructural requirements for connecting the second billion individuals to the Internet. Since then, close to 2 billion people are estimated to have been connected to the Internet, up from the just over one billion at the time of writing (Sanou, 2015). However, it seems increasingly unlikely that the current 'Next Billion' will be connected as easily as the previous ones.

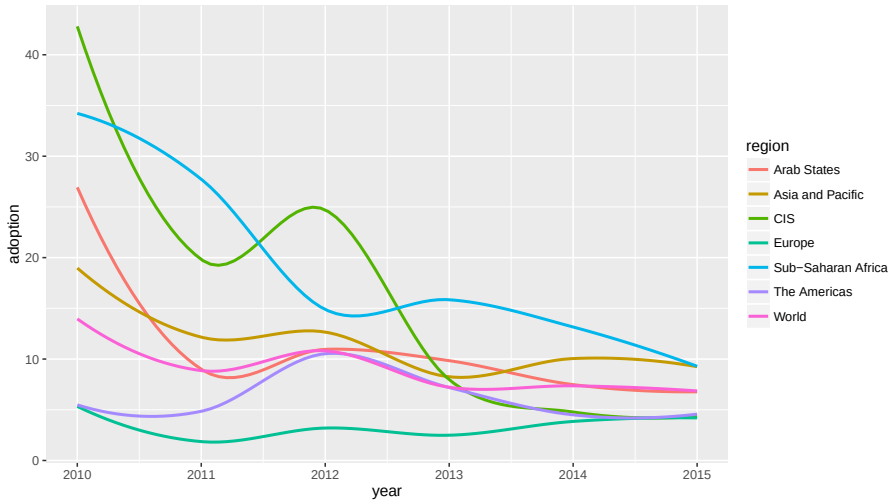
In the period 2010-2014 the average annual growth of Internet bandwidth in sub-Saharan Africa was over fifty percent. This increased bandwidth also causes downward pressure on the cost of Internet access, which brought the sub-Saharan average cost of a 500MB prepaid Internet bundle down to around $10, increasingly putting it within range of the emerging middle classes. Yet, despite increased range and improved affordability, sub-Saharan Africa is showing stagnation in the growth of Internet connected individuals.

As is shown in 2.1, growth of Internet usage in Sub-Saharan Africa is rapidly decreasing. Unlike in other regions in the figure, this observed stagnation is not a consequence of near market saturation, as adoption levels are still relatively low.

A crucial factor in Internet adoption by native speakers of a language is the interplay with content creators using that language. This dynamic is know as a two-sided market, which is characterised as having two different sides, which exhibit positive cross-side network effects (Parker and Van Alstyne, 2000a,b, 2005).

In the case of Internet adoption in a certain language, these sides are on the

Figure 2.1: Internet Adoption



one hand the content creators, such as news websites and on the other hand the content consumers, or Internet users. Ideally, adoption should follow a virtuous circle, whereby the content offering encourages more users to come online, which in turn incentivises more content creation and so on (Rochet and Tirole, 2003, 2006). Unfortunately this virtuous circle sometimes fails to properly start for certain languages, this is especially important as it typically concerns communities who's linguistic characterisation can already hamper economic growth (Arcand and Grin, 2013). Herein also lies the difficulty with finding empirical evidence supporting these dynamics, since the process of adoption by users and content creators is inherently endogenous. With this paper I seek to empirically address the question if increased accessibility of content does indeed lead to an increase in Internet adoption.

Much research has been done on Internet and language with respect to the preservation of smaller languages, in particular indigenous languages. The focus here is often the preservation of the language, such as the below comment by Wikipedia founder Jimmy Wales (Forbes, 2010).

> The Web is a powerful tool for preserving languages that would otherwise be lost. We see this in a lot of the smaller European languages that have very active Wikipedia projects. For example, the Welsh Wikipedia is quite an active community and they have 27,000 articles and this is true even though virtually everyone who speaks Welsh also speaks English.

Unfortunately, this is less true for many indigenous languages outside of Europe. Increasingly language availability is also being considered as a method for im-

proving demand for connectivity (Gandal, 2006; Pena-Lopez, 1999) as well as more generally in development economics as a whole (Arcand, 1996). In a recent paper,Viard and Economides (2014) use macro level connectivity data and a model whereby countries that share languages are used to isolate the effect of content on demand for connectivity, which they find to be positive and significant.

There are good reasons for striving for increased connectivity of remote populations. Jensen and Oster (2009) find that the introduction of cable television in Indian states has a pronounced positive effect on attitudes towards the oppression of women, violence against women, son preference, and as well as decreased fertility. Internet access can provide an in some ways similar window on the outside world, we might expect some of these things to also follow more widespread Internet adoption. Sinai and Waldfogel (2004) show that expanded Internet usage in cities can lead to less racism, an issue that can be of particular relevance to South Africa. The overcoming of ethnic polarisation can in turn be conducive to economic growth (Arcand, Guillaumont et al., 2000).

As mentioned above, Internet connectivity is a two-sided market, which makes it difficult to empirically isolate a causal effect of content availability on demand for Internet connectivity. This paper exploits an exogenous increase in the accessibility of content in the Setswana language in South Africa, in order to isolate the increase of Internet usage among native speakers. In 2010 Google collaborated with a team of Botswanan linguists (Otlogetswe, 2010) to make its Botswanan website (`google.co.bw`) available in the local language: 'Setswana'. In addition to being spoken in Botswana, there is also a sizeable population of Setswana speakers across the border in South Africa, where it is also one of the official state languages. This led to the Setswana-language interface also being introduced on the South-African Google website (`google.co.za`), as a spillover of the translation work originally performed for Google's Botswanan website. This introduction led to a large increase in the number of native Setswana speakers reporting to have spent some amount of money in the past 30 days on Internet access as well as increased computer ownership.

The Google Search interface represents a very small number of words on the Internet and it is not required to use a certain interface language in order to search for content in this language. Yet, the search page is in many cases the first website viewed by users and thereby has a substantial impact on the decision to further engage or not and if they chose to do so, in which language this will be. Asides from being able to understand the interface of the website, having this interface be in a local language also encourages usage of this local language, which in turn reveals more local language content.

In short, we can identify two main channels through which this promotes

increased online engagement, which together constitute the theory of change. Firstly, the ability to read and understand the words of the interface increases the chance that a user continues to use the website and the Internet at large. Secondly, the visibility of local language content increases the likelihood of the user entering search queries in the local language and thereby finding more content in the local language.

The data used for this study comes from the South African National Income Dynamics Survey, provided by Southern Africa Labour and Development Research Unit (2008, 2012, 2013), the data is further discussed in 2.2. After which 2.3 discusses the methods employed in this study, specifically, the discussion of the identification strategy can be found is 2.3.1 and the use of the Difference-in-Differences estimator in 2.3.2. Further, I present the results of the estimation in 2.4. Finally, I conclude in 2.5.

2.2 Data

The National Income Dynamics Study (NIDS) collects data on a representative set of around ten thousand South African households across several time periods. The first wave was gathered in 2008, the second wave in 2010, and the third wave was gathered in 2012 (Southern Africa Labour and Development Research Unit, 2008, 2012, 2013).

In addition to this, in may 2016 a fourth wave of data has also been partially published, which I used to relate the expanded demographic of Setswana speaking Internet users to improvements in employment levels.

The dataset contains an extensive household questionnaire, which contains detailed information on income and expenditure. In particular, it breaks household expenditure down into many forms of food and non-food expenditure, one of which is household expenditure on Internet access in the last 30 days. In addition to this, the household income is calculated and imputed with other income such as home ownership. The individual (adult) questionnaires also contain information on linguistic skills in both English and in the interviewees native language, as well as a series of variables relating to communication technology ownership and utilisation, in particular computer ownership. In table 2.4 an overview of the dependent variables, broken down by wave and native language is presented.

Table 2.1: Dependent Variable Descriptive Statistics

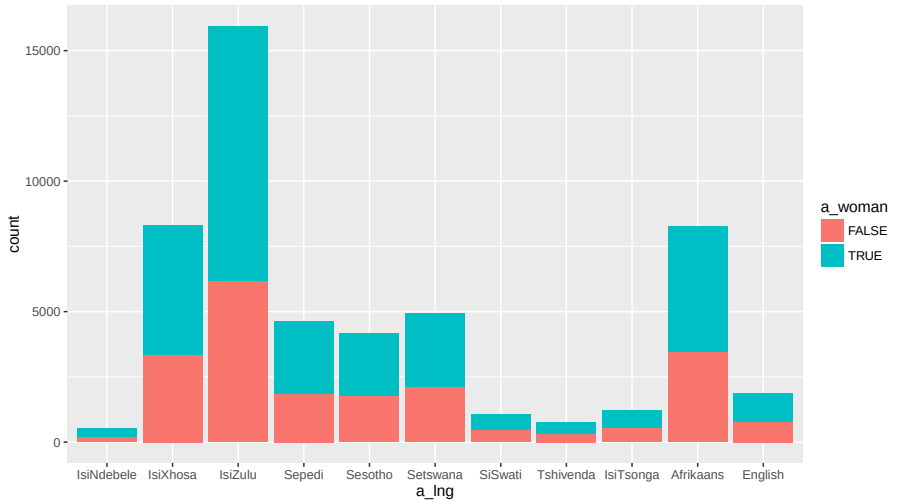
	Wave	Setswana	Other
Computer Ownership	1	0.0351	0.0567
	2	0.0359	0.0417
	3	0.0711	0.0632
Household Internet Expenditure	1	0.0068	0.0162
	2	0.0037	0.0224
	3	0.0106	0.0140

I use both household expenditure on Internet access and computer ownership as dependent variables. Household expenditure includes a variety of way in which this expense can be made, including paying for a fixed-line subscription, a mobile Internet subscription, as well using Internet in an Internet cafe. More than 99% of South Africa is covered by mobile Internet networks, this figure did not change during the time periods used in this study (Union, 2015). Computer ownership is recorded in the individual adult questionnaire, which provides me with more observations, unlike expenditure in the last 30 days, this shows a more long-term investment.

In addition to the variables of interest, I include a number of relevant covariates, such as household income and education levels. Furthermore, I include information on English and native language reading and writing skills. Around 45% of the individuals report being able to read and write English fluently, whereas around 55% report being able to do so in their native language.

In 2.2 we can see the number of native speakers for each language in the dataset, coloured by gender. The dataset contains a total of 51,612 observations (adult individuals), of which 2,806 are female native Setswana speakers and 2,140 are male native Setswana speakers.

Figure 2.2: Native Language and Gender



2.3 Empirical Methodology

With this paper I aim to answer the question whether increased content or accessibility of content leads to an increase in demand. This section begins with a discussion of the identification strategy employed, followed by a explanation of the estimator used to operationalise this.

2.3.1 Identification Strategy

This paper exploits the introduction of the Setswana interface language to Google Search in South Africa as a spillover of the development of that interface for the Botswanan Google Search website. By comparing the number of native Setswana speakers in South Africa being Internet users, with the number of South Africans with a different native language around the same time, I isolate the effect of this introduction.

The Setswana language interface was first developed for the Botswanan Google Search website (`google.co.bw`). As such, the introduction of Setswana to the South African Google Search (`google.co.za`) was a spillover effect of that development. This allows me to rule out any possible endogeneity issues that might otherwise arise in contexts such as these. For instance, the Afrikaans language is almost solely spoken in South Africa. When we observe that the introduction of the Afrikaans Google Search interface occurs around the same time as a growth in the number of native Afrikaans Internet users, it will be hard to isolate the effect from the introduction from its cause (since an increase

in native Afrikaans Internet users would be a good reason to introduce it as an interface language).

Substantial numbers of Setswana speakers exist in Botswana, South Africa, Zimbabwe, and to some extend Namibia. However, the language is most important in Botswana, where it is spoken by approximately 80% of all people, and where it is the only official language other than English. As such, it is also the place where most linguistic work on the Setswana language takes place. The Setswana Google Search interface was also developed at the University of Botswana by prof. Otlogetswe.

It is worth noting that it is very common to not personally own a computer, therefore 'paying for Internet access' also includes a lot of people who use the Internet in other locations such as Internet cafe's.

In addition to using the propensity to spend on Internet (in the last thirty days), I also use the propensity to own a computer as a dependent variable.

2.3.2 Regression Specification

As mentioned in the above section, I compare the change in the level of Internet users among native Setswana speakers in South Africa, with that of native speakers of other language in South Africa around the introduction of the Setswana interface to the South-African Google Search, after the second wave of the NIDS. For this I use a Difference-in-Differences estimator (Abadie, 2005; Imbens and Jeffrey M. Wooldridge, 2009) using a native-Setswana speaker dummy variable (`setswana`), interacted with an event dummy variable (`event`). The former is `TRUE` when the native language of the individual (`a_lng`) is `Setswana` and `FALSE` otherwise. The latter is `FALSE` for data collected prior to the introduction of the Setswana interface language (late 2010, here wave 1 and 2) and `TRUE` after this introduction (here wave 3). The model than takes the form as described in equation (2.1).

$$y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \beta X_{it} + \epsilon_{it} \tag{2.1}$$

Where α_i represents the individual fixed effects, λ_t represent the time fixed effects, and X_{it} are the time varying covariates. The ϵ_{it} is the error term. Finally the term of interest is D_{it} which represents the treatment effect.

The `h_nfnet` variable is recorded at a household level, as such a standard error correction needs to be applied (White, 1980), the model with standard error corrections is reported in the appendix.

Lastly, the dependent variables are both logical or binary variables, as such, normally a model such as logit should be used. However, since I am using Difference-in-Differences, this model would be undefined (Jeffrey M Wooldridge,

2010), I therefore use a standard linear model.

2.4 Results

In the base model, I use an interaction of the `event` dummy and `setswana` dummy in order to isolate the effect on the explanandum, a dummy variable describing household expenditure on Internet in the last thirty days or not (`Internet_expenditure`, household non-food Internet). The results of this estimation are presented in table 2.2.

I find that the interaction term of the event dummy (`event`) and the native Setswana speaker dummy (`setswana`) is positive and highly significant, with a p-value around `0.0018`. Both the individual dummy variables (`event` and `setswana`) yield significant negative parameter estimates.

In addition to this, the covariates included in the estimation are also highly significant. The highest education level of the individual (`best_edu`) and the household income (`hhincome`) are both positive and significant. The parameter estimate of `woman` here is negative but not at all significant, this is unsurprising as I use Internet expenditure at a household level. Most women live in a household which includes men and visa versa, suggesting that this effect cannot be isolated in this estimation. I further investigate this issue in a separate estimation discussed below. The variables describing linguistic skills in reading and writing in both English and the native language do yield many significant results, though lower levels of English writing skill seems to be correlated with a lower propensity to use the Internet (`a_edlitwrten` for levels `2` and `3`, but not the very lowest: `4`).

In an alternative formulation, I include the native language variable as a categorical variable (`language`), interacted with the `event` dummy. In this estimation I only find significantly positive results for `Setswana` and `Venda` (as small language from the region bordering Zimbabwe), and a significantly negative effect for the language `Afrikaans`.

When using the propensity of adults (`own_computer`) to own a computer is used as an explanandum, I find similar results. This is of particular relevance, as the explanandum here (`own_computer`) differs from the base model's explanandum in two ways. Firstly, it does not include expenditure on Internet in ways such as Internet cafes, but focusses on actual ownership, signalling a more long-term investment and interest. Secondly, the `Internet_expenditure` variable is at a household level, whereas the `own_computer` variable is at the level of an individual adult. The results from this estimation are included in table 2.2. This form of the estimation yields similar results to those estimated in the base model. Firstly I find that the variable of interest, the interaction term

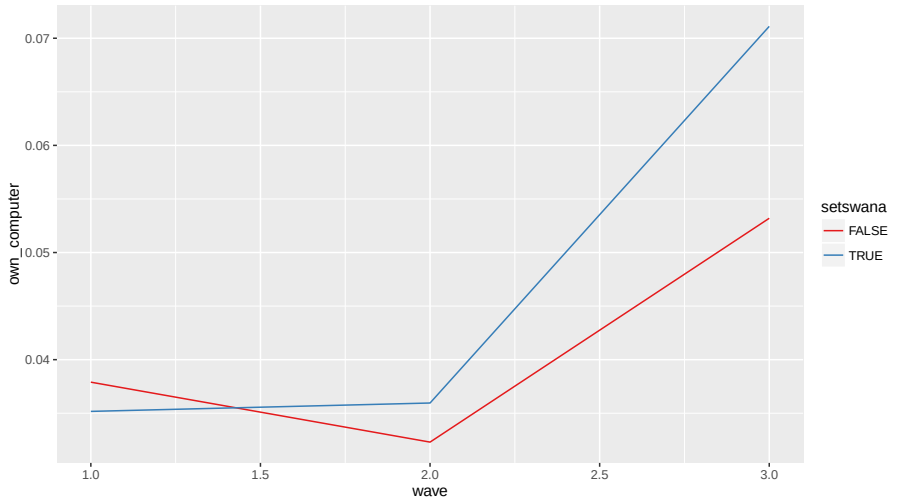
Table 2.2: Internet Access and Computer Ownership

	Internet	(P > |t|)	Computer	(P > |t|)
event * setswana	**0.012**	0.00	**0.024**	0.00
event	**-0.012**	0.00	-0.054	0.01
setswana	**-0.014**	0.00	**-0.015**	0.00
income	**0.000**	0.00	**0.000**	0.00
woman	-0.001	0.23	**-0.023**	0.00
education	**0.001**	0.00	**0.006**	0.00
Observations	47665		46464	

between the event and the language dummy (`event * setswana`) is positive at 0.024 and highly significant, with a p-value smaller than `0.001`. This means that the event increased the propensity to own a computer by 2.4%. The individual dummy variables (`event` and `setswana`) again are significant and negative with the former's p-value smaller than `0.01` and the latter's smaller than `0.001`. In terms of the linguistic skill, I find that the lower levels of English reading as well as English writing are correlated with lower propensities of computer ownership. Similar to Internet expenditure model, household income (`hhincome`) and highest level of education (`best_edu`) are both positive and highly significant (p-value: `~0`). However, unlike in the household Internet expenditure model, the gender of the individual here is highly significant, specifically, parameter estimate of `woman` is negative and highly significant (p-value: `~0`). As mentioned above, this variable is difficult to interpret when using a household-level variable as an explanandum, however, here, the computer ownership variable is at an individual level, which makes the coefficient more easily interpretable.

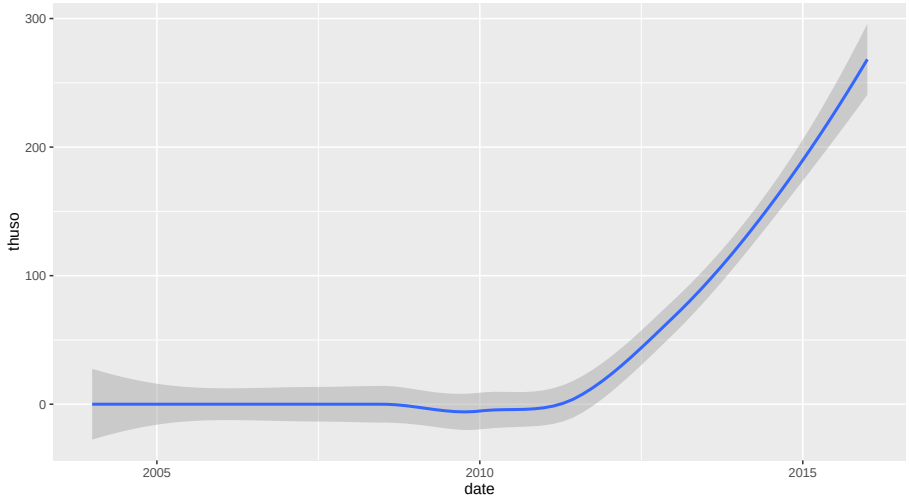
The below figure present a graphical illustration of the above mentioned result. As we can see, initially, the propensity to own a computer for setswana speakers (blue line) was similar to the average of speakers of other languages (red line), however after the introduction of the Setswana language search interface, between waves 2 and 3, we observe a sharp increase in this propensity for setswana speakers in wave 3, whereas average of other languages remains relatively unchanged.

Figure 2.3: Computer Ownership Setswana



2.4, illustrates how usage of the Setswana word 'thuso', meaning 'help' in search queries, was zero prior to the introduction of the Setswana search interface and became widespread thereafter. Therefore, in addition to the increased Internet usage by Setswana speakers, we can also observe an increased usage of the Setswana language itself. This can lead to greater amounts of Setswana language content being found and engaged with, which in turn incentivised content creators to provide more Setswana language content. This could help break the vicious circle of low levels of content and few users.

Figure 2.4: Usage of Setswana Words on Google.co.za

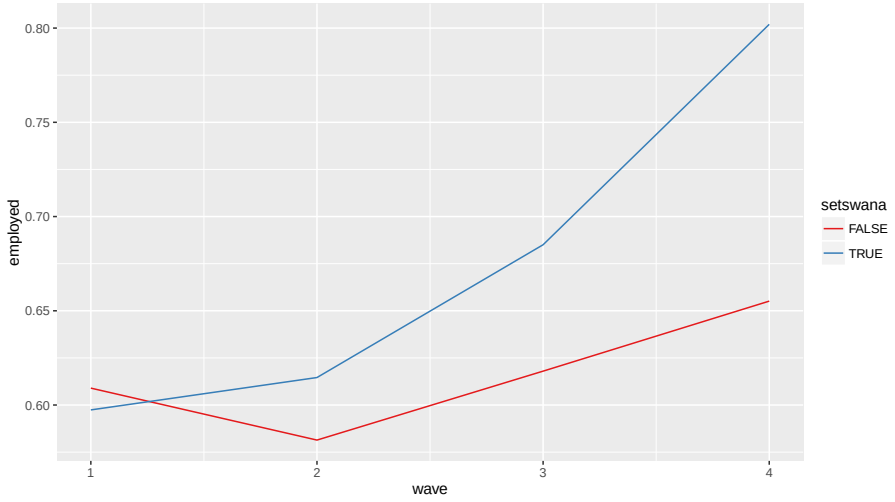


As mentioned in section §2.2 I use the partially released 4th wave of the

dataset, which covers the year 2014 in order to relate the increased Internet usage with employment levels. I do this by subsetting data to include one the one hand, only the individuals that owned a computer in after the event in wave 3 and on the other hand the individuals in households that reported spending on internet access. The idea is that the owners/users in wave 3 are an expanded demographic for Setswana speakers but remain relatively unchanged. With this, we can see the evolution of the employment level for this expanded demographic.

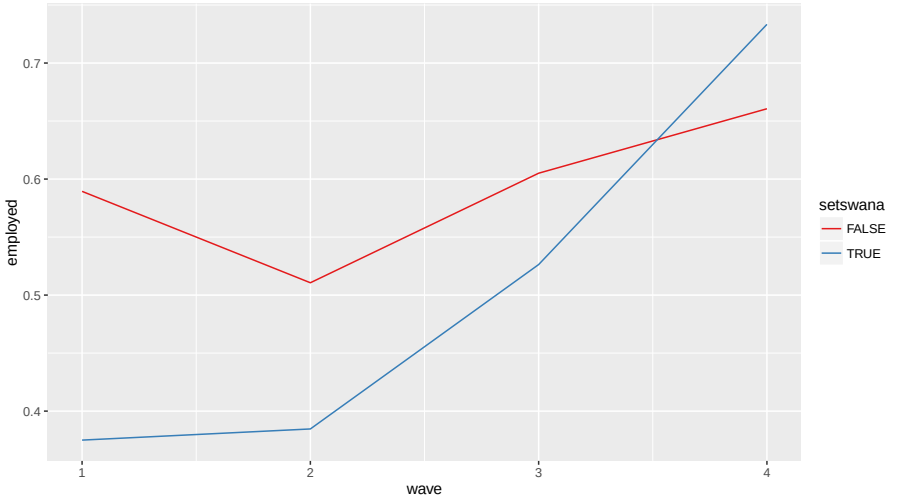
In the below figure, I plot the employment status of individuals that own a computer in the first wave after the introduction of the Setswana-language interface (wave 3). The figure shows that there is a sharp uptick in the proportion of employed individuals among Setswana speakers.

Figure 2.5: Employment for Individuals who Own a Computer in Wave 3



Similarly, the individuals that lived in a household that spent on Internet access in the last 30 days also saw a strong increase in the proportion of employed individuals, overtaking the proportion for the rest of the population.

Figure 2.6: Employment for Individuals with Internet Expenditure in Wave 3



2.5 Conclusions and Limitations

In conclusion, despite recent advances in the reach, speed, and affordability of Internet connectivity in sub-Saharan Africa, actual uptake has been stagnant. Internet adoption is a two-sided market, as a result, one aspect of this is that users benefit from cross-side network effects from content. Unfortunately, in certain languages, a sufficient amount of content is not always created, as a result, Internet adoption among native speakers of these languages can lag behind.

This paper demonstrates that this failure can in part be attributed to the dynamics of a two-sided market, whereby the vicious circle of few users and little content perpetuates a situation of low levels of adoption. Due to the endogenous nature of two-sided markets, there are few methods of isolating a causal effect.

I exploit the introduction of the Setswana-language interface on `google.co.za`, as a spillover of the development of this interface for `google.co.bw`, I find that it leads to a substantial increase in Internet usage and computer ownership among native Setswana speakers.

Furthermore, when comparing the Setswana speakers that own a computer or spend on Internet access after the event, with non Setswana speakers, we see a marked increase in the proportion of employed individuals over that among the rest of the population. This increase is even stronger in wave 4 of the data, suggesting that it takes some time for the effect to fully materialise and that it is persistent.

This increase of Internet usage among the Setswana speaking population as a result of the newly introduced interface language on `google.co.za`, suggests

that there is a serious lack in the availability of local content in many indigenous African languages, which serves as an impediment to further Internet adoption.

This suggests that the effect is unlikely to be ephemeral in nature, since computer ownership constitutes a more long-term investment in Internet access.

As I discussed in the results section, the increase in computer ownership for Setswana speakers between wave 2 and wave 3, was 115%, compared to 70% for the rest of the population, or a 50% greater increase. The effect on Internet expenditure in the household, which includes expenditure in Internet cafes etc. is even greater. The proportion of Setswana households that spent on Internet in the last 30 days increased by 217%, whereas for the rest of the population it fell be 22%.

Bibliography

Abadie, Alberto

2005 "Semiparametric difference-in-differences estimators", *The Review of Economic Studies*, 72, 1, pp. 1–19.

Arcand, Jean-Louis

1996 "Development economics and language: the earnest search for a mirage?", *International Journal of the Sociology of Language*, 121, 1, 243Ã¢ÂÂ266.

Arcand, Jean-Louis and Francois Grin

2013 "Language in Economic Development: Is English Special and is Linguistic Fragmentation Bad?", in *English and development: Policy, pedagogy and globalization*, ed. by Elizabeth J. Erling, Multilingual Matters, chap. 11, pp. 243–266.

Arcand, Jean-Louis, Patrick Guillaumont, Sylviane Guillaumont Jeanneney et al.

2000 *Ethnicity, communication and growth*, University of Oxford, Department of Economics, Centre for the Study of African Economies.

Croissant, Yves

2013 "pglm: Panel Generalized Linear Model", *R package version 0.1-2*, `http://CRAN.%20R-project.%20org/package=%20pglm`.

Forbes

2010 "Jimmy Wales: The Wiki World", *Forbes*, `http://www.forbes.com/2010/06/15/forbes-india-jimmy-wales-the-wiki-world-opinions-ideas-10-wales.html`.

Gandal, Neil

2006 "Native language and Internet usage", *International journal of the sociology of language*, 2006, 182, pp. 25–40.

Git Team

2016 *Git: Software Code Manager*, 137 Montague ST STE 380, Brooklyn, NY 11201-3548, `http://www.git-scm.org/`.

Hoekwater, Taco, Hartmut Henkel and Hans Hagen

2016 *LuaTeX*, `http://www.luatex.org/`.

Imbens, Guido W. and Jeffrey M. Wooldridge

2009 "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, 47, 1 (Mar. 2009), pp. 5–86, DOI: `10.1257/jel.47.1.5`, `http://www.aeaweb.org/articles/?doi=10.1257/jel.47.1.5`.

Jensen, Robert and Emily Oster

2009 "The power of TV: Cable television and women's status in India", *The Quarterly Journal of Economics*, pp. 1057–1094.

Knuth, Donald Ervin

1984 "Literate programming", *The Computer Journal*, 27, 2, pp. 97–111.

Lamport, Leslie

1985 *I1 (\ LaTeX)—A Document*, pub-AW, vol. 410.

Leisch, Friedrich

2002 "Sweave: Dynamic generation of statistical reports using literate data analysis", in *Compstat*, Springer, pp. 575–580.

LyX Team

2016 *LyX*, Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA, `http://www.lyx.org/`.

Otlogetswe, Thapelo J.

2010 "Setswana Google is here!", *T.J. Otlogetswe Blog*, `http://otlogetswe.com/2010/08/13/setswana-google-here/`.

Parker, Geoffrey G and Marshall W Van Alstyne

2000a "Information complements, substitutes, and strategic product design", in *Proceedings of the twenty first international conference on Information systems*, Association for Information Systems, pp. 13–15.

2000b "Internetwork externalities and free information goods", in *Proceedings of the 2nd ACM Conference on Electronic Commerce*, ACM, pp. 107–116.

Parker, Geoffrey G and Marshall W Van Alstyne

2005 "Two-sided network effects: A theory of information product design", *Management science*, 51, 10, pp. 1494–1504.

Pena-Lopez, Ismael

1999 "Challenges to the Network: Internet for Development".

R Core Team

2016 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`.

Rochet, Jean-Charles and Jean Tirole

2003 "Platform competition in two-sided markets", *Journal of the European Economic Association*, 1, 4, pp. 990–1029.

2006 "Two-sided markets: a progress report", *The RAND journal of economics*, 37, 3, pp. 645–667.

Sanou, Brahima

2015 "The World in 2015: ICT facts and figures", *International Telecommunications Union*.

Sinai, Todd and Joel Waldfogel

2004 "Geography and the Internet: Is the Internet a Substitute or a Complement for Cities?", *Journal of Urban Economics*, 56, 1, pp. 1–24.

Southern Africa Labour and Development Research Unit

2008 *National Income Dynamics Study, Wave 1*, version 5.1, `http://www.nids.uct.ac.za/home/`.

2012 *National Income Dynamics Study, Wave 2*, version 2.1, `http://www.nids.uct.ac.za/home/`.

2013 *National Income Dynamics Study, Wave 3*, version 1.1, `http://www.nids.uct.ac.za/home/`.

Standage, Tom

2006 "Connecting the next billion", *The Economist-The World in 2006*, p. 117, `http://www.economist.com/node/5134746`.

Union, International Telecommunication

2015 *The World in 2011: ICT Facts and Figures*, ITU.

Venables, William N and Brain D. Ripley

2013 *Modern applied statistics with S-PLUS*, Springer Science & Business Media.

Viard, V Brian and Nicholas Economides

2014 "The Effect of Content on Global Internet Adoption and the Global Ã¢ÂÂDigital DivideÃ¢ÂÂ", *Management Science*, 61, 3, pp. 665–687.

White, Halbert

1980 "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", *Econometrica: Journal of the Econometric Society*, pp. 817–838.

Wooldridge, Jeffrey M

2010 *Econometric analysis of cross section and panel data*, MIT press.

Xie, Yihui

2015 *Dynamic Documents with R and knitr*, Chapman and Hall/CRC, vol. 29, ISBN: 978-1498716963, `http://yihui.name/knitr/`.

Zeileis, Achim

2004 "Econometric Computing with HC and HAC Covariance Matrix Estimators", *Journal of Statistical Software*, 11, 10, pp. 1–17, `http://www.jstatsoft.org/v11/i10/`.

2006 "Object-Oriented Computation of Sandwich Estimators", *Journal of Statistical Software*, 16, 9, pp. 1–16, `http://www.jstatsoft.org/v16/i09/`.

Zeileis, Achim and Torsten Hothorn

2002 "Diagnostic Checking in Regression Relationships", *R News*, 2, 3, pp. 7–10, `http://CRAN.R-project.org/doc/Rnews/`.

2.A Clustering

Table 2.3: Clustering

```
library(plm)     # panel linear model estimation
library(lmtest) # Standard Error corrections
library(broom)  # output formatting using tidy()

# specify panel model
plm4_3 <- formula(as.numeric(h_nfnet)  ~ interface_intro*setswana  +
                                         factor(a_edlitrden)  +
                                         factor(a_edlitwrten) +
                                         factor(a_edlitrdhm)  +
                                         factor(a_edlitwrthm) +
                                         a_woman              +
                                         hhincome             +
                                         best_edu              )
# estimate
plm4_3e <- plm(plm4_3, data=pNIDS, model='within')

# correct errors
tidy( coeftest(plm4_3e, vcov=vcovHC(plm4_3e,
                                    type="HC0",
                                    cluster="group")) )
```

term	estimate	std.error	statistic	p.value
interface_introTRUE	-0.0019402	0.0012392	-1.5657830	0.1174144
setswanaTRUE	0.0027958	0.0137038	0.2040204	0.8383396
factor(a_edlitrden)2	0.0061294	0.0039180	1.5643993	0.1177388
factor(a_edlitrden)3	0.0028083	0.0042618	0.6589578	0.5099301
factor(a_edlitrden)4	0.0001637	0.0044802	0.0365363	0.9708551
factor(a_edlitwrten)2	-0.0088583	0.0039162	-2.2619681	0.0237095
factor(a_edlitwrten)3	-0.0080226	0.0042172	-1.9023800	0.0571351
factor(a_edlitwrten)4	-0.0060115	0.0046454	-1.2940715	0.1956549
factor(a_edlitrdhm)2	-0.0029617	0.0043786	-0.6764141	0.4987852
factor(a_edlitrdhm)3	-0.0035411	0.0061579	-0.5750552	0.5652601
factor(a_edlitrdhm)4	-0.0056185	0.0068075	-0.8253295	0.4091939
factor(a_edlitwrthm)2	0.0020764	0.0043717	0.4749531	0.6348253
factor(a_edlitwrthm)3	0.0045419	0.0063446	0.7158756	0.4740761
factor(a_edlitwrthm)4	0.0077089	0.0074611	1.0332094	0.3015178
a_womanTRUE	-0.0012994	0.0012422	-1.0461001	0.2955268
hhincome	0.0000007	0.0000003	2.4693445	0.0135439
best_edu	0.0001411	0.0003131	0.4507013	0.6522095
interface_introTRUE:setswanaTRUE	0.0070425	0.0034066	2.0673056	0.0387175

2.B Dependent Variable Breakdown

Table 2.4: Descriptive statistics on Ownership and Expenditure

a_lng	wave	a_owncom	h_nfnet
IsiNdebele	1	0.0331126	0.0000000
IsiNdebele	2	0.0270270	0.0284091
IsiNdebele	3	0.0333333	0.0000000
IsiXhosa	1	0.0112269	0.0012043
IsiXhosa	2	0.0186335	0.0054517
IsiXhosa	3	0.0345508	0.0019756
IsiZulu	1	0.0132693	0.0013730
IsiZulu	2	0.0127744	0.0086366
IsiZulu	3	0.0252420	0.0044928
Sepedi	1	0.0265273	0.0048309
Sepedi	2	0.0226818	0.0021994
Sepedi	3	0.0718697	0.0050477
Sesotho	1	0.0366044	0.0062598
Sesotho	2	0.0457010	0.0213640
Sesotho	3	0.0949535	0.0113032
Setswana	1	0.0351724	0.0068681
Setswana	2	0.0359537	0.0037783
Setswana	3	0.0711086	0.0106264
SiSwati	1	0.0441640	0.0000000
SiSwati	2	0.0612813	0.0091185
SiSwati	3	0.0458221	0.0134771
Tshivenda	1	0.0334928	0.0000000
Tshivenda	2	0.0000000	0.5441176
Tshivenda	3	0.0225080	0.0000000
IsiTsonga	1	0.0235294	0.0000000
IsiTsonga	2	0.0118203	0.0929368
IsiTsonga	3	0.0397196	0.0023529
Afrikaans	1	0.1345441	0.0465116
Afrikaans	2	0.0904605	0.0424710
Afrikaans	3	0.1106225	0.0399458
English	1	0.2969374	0.1016043
English	2	0.3234127	0.1070707
English	3	0.3156934	0.1023766

2.C Covariate Descriptive Statistics

Figure 2.7: Household Income

```
ggplot(adulthh, aes(x=hhincome, fill=a_lng )) +
    stat_bin(bins=50)
```

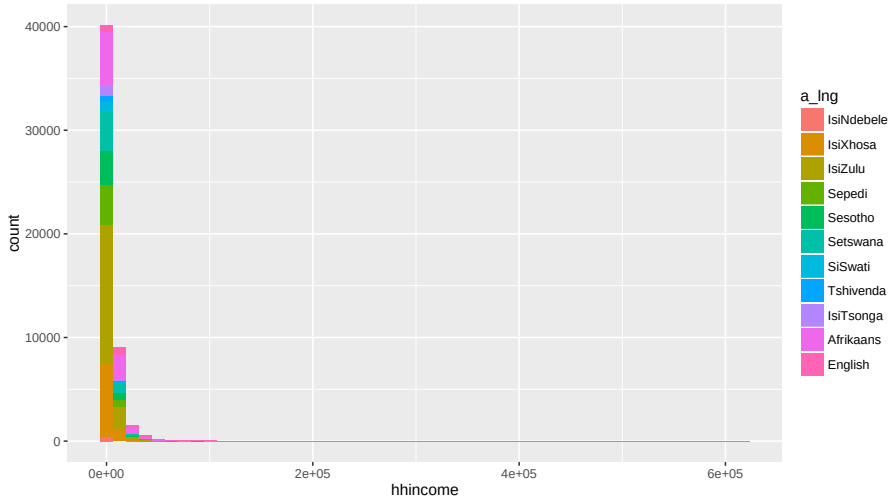
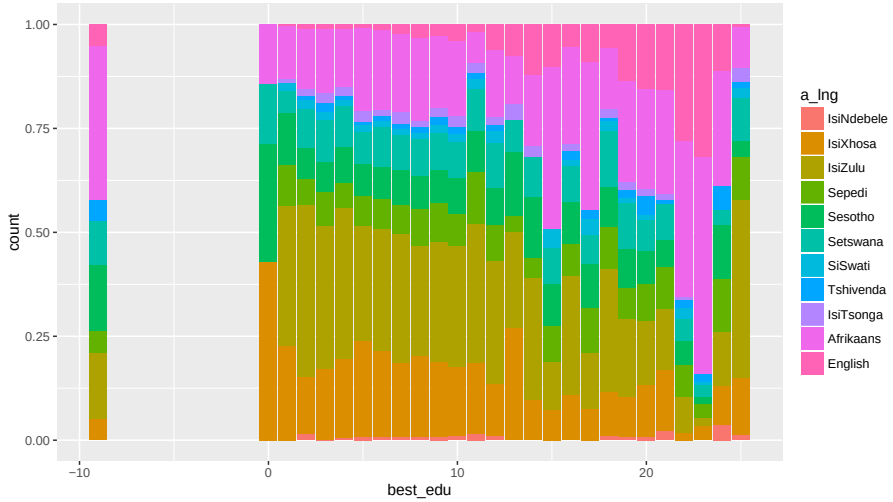


Figure 2.8: Years of Education

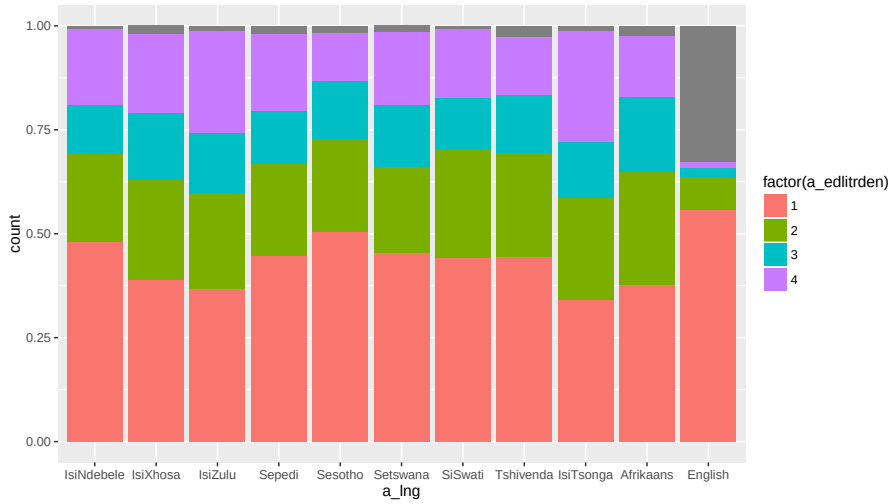
```
ggplot(adulthh, aes(x=best_edu, fill=a_lng )) +
    geom_bar(position = 'fill')
```



The 2.9 describes the skill of individuals in reading the English language, where 1 the best and 4 is the worst, grey values are `NA`.

Figure 2.9: English Language Reading Skills

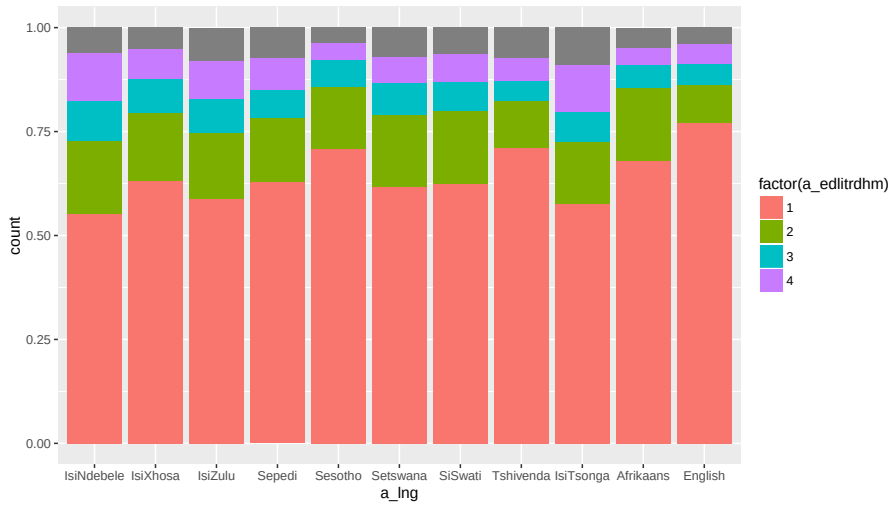
```
ggplot(adulthh, aes(x = a_lng, fill = factor(a_edlitrden)) ) +
        geom_bar(position = 'fill')
```



The 2.10 does the same but with regards to the native language.

Figure 2.10: Native Language Reading Skills

```
ggplot(adulthh, aes(x = a_lng, fill = factor(a_edlitrdhm)) ) +
        geom_bar(position = 'fill')
```



2.D Original Estimates

Table 2.5: Computer Ownership

```
lm(a_owncom ~ interface_intro*setswana_logical +
               factor(a_edlitrden)                          +
               factor(a_edlitwrten)                         +
               factor(a_edlitrdhm)                          +
               factor(a_edlitwrthm)                         +
               a_woman                                      +
               hhincome                                     +
               best_edu)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0075393	0.0031352	2.4047103	0.0161891
interface_introTRUE	-0.0053820	0.0020917	-2.5729824	0.0100856
setswana_logicalTRUE	-0.0147502	0.0041653	-3.5411916	0.0003987
factor(a_edlitrden)2	-0.0306749	0.0069077	-4.4406931	0.0000090
factor(a_edlitrden)3	-0.0310090	0.0090330	-3.4328617	0.0005978
factor(a_edlitrden)4	-0.0389174	0.0116674	-3.3355679	0.0008519
factor(a_edlitwrten)2	-0.0173238	0.0068925	-2.5134122	0.0119602
factor(a_edlitwrten)3	-0.0210640	0.0089384	-2.3565895	0.0184476
factor(a_edlitwrten)4	-0.0187291	0.0114540	-1.6351572	0.1020227
factor(a_edlitrdhm)2	-0.0017925	0.0063219	-0.2835349	0.7767681
factor(a_edlitrdhm)3	-0.0041655	0.0088001	-0.4733526	0.6359638
factor(a_edlitrdhm)4	-0.0267964	0.0118797	-2.2556443	0.0240974
factor(a_edlitwrthm)2	0.0010896	0.0063817	0.1707311	0.8644360
factor(a_edlitwrthm)3	-0.0020020	0.0088176	-0.2270503	0.8203856
factor(a_edlitwrthm)4	-0.0357886	0.0118921	-3.0094517	0.0026186
a_womanTRUE	-0.0230335	0.0019598	-11.7530143	0.0000000
hhincome	0.0000058	0.0000001	56.8794124	0.0000000
best_edu	0.0058419	0.0002002	29.1761771	0.0000000
interface_introTRUE:setswana_logicalTRUE	0.0238052	0.0066799	3.5637169	0.0003660

Table 2.6: Living in Household that Spent on Internet (last 30 days)

```
lm(h_nfnet ~  interface_intro*setswana_logical +
              factor(a_edlitrden)                          +
              factor(a_edlitwrten)                         +
              factor(a_edlitrdhm)                          +
              factor(a_edlitwrthm)                         +
              a_woman                                      +
              hhincome                                     +
              best_edu)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0008565	0.0018351	-0.4667340	0.6406924
interface_introTRUE	-0.0117376	0.0012170	-9.6449716	0.0000000
setswana_logicalTRUE	-0.0135286	0.0024255	-5.5777454	0.0000000
factor(a_edlitrden)2	0.0015879	0.0040272	0.3942904	0.6933684
factor(a_edlitrden)3	0.0002543	0.0052755	0.0481963	0.9615600
factor(a_edlitrden)4	-0.0036574	0.0068160	-0.5365861	0.5915561
factor(a_edlitwrten)2	-0.0101933	0.0040178	-2.5369998	0.0111839
factor(a_edlitwrten)3	-0.0107676	0.0052214	-2.0621916	0.0391951
factor(a_edlitwrten)4	-0.0071685	0.0066937	-1.0709421	0.2842010
factor(a_edlitrdhm)2	-0.0025418	0.0036848	-0.6898084	0.4903181
factor(a_edlitrdhm)3	-0.0028899	0.0051291	-0.5634293	0.5731453
factor(a_edlitrdhm)4	-0.0079926	0.0069070	-1.1571655	0.2472107
factor(a_edlitwrthm)2	0.0008117	0.0037229	0.2180373	0.8274010
factor(a_edlitwrthm)3	0.0013993	0.0051372	0.2723915	0.7853222
factor(a_edlitwrthm)4	-0.0069174	0.0069137	-1.0005354	0.3170567
a_womanTRUE	-0.0013881	0.0011452	-1.2121278	0.2254696
hhincome	0.0000027	0.0000001	46.2987087	0.0000000
best_edu	0.0013582	0.0001164	11.6710813	0.0000000
interface_introTRUE:setswana_logicalTRUE	0.0119717	0.0038666	3.0961443	0.0019617

2.E Software

The estimation is primarily performed using R (R Core Team, 2016), specifically using `lm()` and `glm()` functions included in the `stats` package (Venables and Ripley, 2013). Additionally, I make the of the `plm()` and `pglm()` functions which are available in packages by the same names (**croissant2008plm**; Croissant, 2013). Standard error corrections are computed using the `lmtest` and `sandwich` packages Zeileis (2004, 2006); Zeileis and Hothorn (2002).

In order to make the result as easily reproducible as possible, this research and writing in the article has been done exclusively using open-source software such as R (R Core Team, 2016). This document is written and LyX (LyX Team, 2016) in the LaTeX(Lamport, 1985) language and compiled using the LuaTeX implementation(Hoekwater et al., 2016). The integration of R code and output in the document is performed using a process call literate programming Knuth (1984) using the knitr implementation Xie (2015) of the Sweave framework (Leisch, 2002).

All changes are logged using the version control system Git (Git Team, 2016) and publicly available on GitHub at https://github.com/bquast/Making-Next-Billion-Demand-Access/[1]

[1]The repository can by cloned to a local computer by entering in following command in a terminal (with Git installed):
`git clone https://github.com/bquast/Making-Next-Billion-Demand-Access.git`

Chapter 3

`decompr`: Global Value Chain decomposition in `R`

with Victor Kummritz

Abstract

Global Value Chains have become a central unit of analysis in research on international trade. However, the complex matrix transformations at the basis of most Value Chain indicators still constitute a significant entry barrier to the field. The R package decompr solves this problem by implementing the algorithms for the analysis of Global Value Chains as R procedures, thereby simplifying the decomposition process. Two methods for gross export flow decomposition using Inter-Country Input-Output tables are provided. The first method concerns a decomposition based on the classical Leontief (1936) insight. It derives the value added origins of an industry's exports by source country and source industry, using easily vailable gross trade data. The second method is the Wang-Wei-Zhu algorithm which splits bilateral gross exports into 16 value added components. These components can broadly be divided into domestic and foreign value added in exports. Using the results of the two decompositions, decompr provides a set of Global Value Chain indicators, such as the now standard Vertical Specialisation ratio. This article summarises the methodology of the algorithms, describes the format of the input and output data, and exemplifies the usefulness of the two methods on the basis of a simple example data set.

3.1 Introduction

Global Value Chains (GVCs) refer to the quickly expanding internationalization of production networks. Most goods we use nowadays consist of parts that are sourced from different corners of the planet and are assembled across different continents. A popular example of this development is the iPhone, which uses inputs from at least five countries (USA, China, Germany, Taiwan, South Korea) and is assembled in two (USA and China). This has made GVCs a central topic in research on trade and development policy. Both policy makers and academia increasingly value the growth opportunities GVCs offer to global trade and, especially, to developing countries. However, analysing this phenomenon empirically requires complex matrix manipulations, since the relevant data is only available in the form of gross flows. The decompr package enables researchers with little background in matrix algebra and linear programming to easily derive standard GVC indicators for statistical analysis.

The package uses Inter-Country Input-Output tables (ICIOs), such as those published by the OECD and WTO (TiVA), the World Input Output Database (Timmer et al., 2012), or national statistics bureaus, as input. These tables state supply and demand relationships in gross terms between industries within and across countries. For instance, let us look at the example of the leather used in German manufactured car seats. The ICIOs quantify the value of inputs that the Turkish leather and textiles industry supplies to the German transport equipment industry. The problem of these tables measuring gross trade flows, is that they do not reveal how much of the value was added in the supplying industry, and how much of the value was added in previous stages of production, performed by other industries or even countries.

The Leontief decomposition of gross trade flows solves this problem by reallocating the value of intermediate goods used by industries to the original producers. In our example, the use of Argentinian agricultural produce (raw hides) is subtracted from the Turkish leather industry and added to the Argentinian agricultural industry. The Wang-Wei-Zhu (henceforth WWZ) decomposition goes a step further by not only revealing the source of the value added, but also breaking down exports into different categories, according to final usage and destination. It implements the theoretical work of Wang et al. (2014). The main categories in this framework are listed below.

1. domestic value added in exports
2. foreign value added in exports
3. pure double counted terms

3.1.1 Package Details

The decompr package implements the algorithms for these decompositions as R procedures and provides example data sets. We start by loading the package and listing the functions.

```
# load package
library(decompr)

# list functions in package
ls('package:decompr')

## [1] "decomp"              "leontief"            "load_tables"
## [4] "load_tables_vectors" "wwz"
```

The R procedures are implemented as functions, the included functions are listed below.

- `load_tables_vectors()`; transforms the input objects to an object used for the decompositions (class: `decompr`)
- `leontief()`; takes a decompr object and applies the Leontief decomposition
- `wwz()`; takes a decompr object and applies the Wang-Wei-Zhu decomposition
- `decomp()`; a wrapper function which integrates the use of load_tables_vectors with the various decompositions, using an argument *method* to specify the desired decomposition (default leontief)

For legacy purposes, one depracated function is also available under their original names (`load_tables()`). In addition to this, one example data sets is included.

- `leather`; a fictional three-country, three-sector data set [1]

Trade flow analysis often involves studying the development of a certain variable (set) over time, thus taking the panel form. However, at the decomposition level, the panel dimension is essentially a repeated cross-section. Therefore, as a design decision, the time dimension is not implemented in the package itself. Instead, we provide examples of how this repetition can be implemented using a `for-loop`.

[1] load using: `data(leather)`

section §3.2 introduces the data as it is used by the package as well as two example data sets, after which section §3.3 and section §3.4 summarise the theoretical derivations for the two decompositions, and show how these can be performed in R using decompr. We conclude with a discussion of potential uses and further developments of GVC research.

3.2 Data

Two data sets are included in the package, one real world data set and one minimal data set for demonstration purposes. The former is the WIOD regional Inter-Country Input Output tables for the year 2011 (Timmer et al., 2012). The latter is a fictional 3-country 3-sector data set, which we will use throughout this article to demonstrate the usage and advantages of the decompr package.

```
# load the data
data(leather)

# list the objects in the data set
ls()
```

```
##  [1] "ITU2015"     "NIDS"        "adulthh"     "child"       "countries"
##  [6] "df"          "expf1"       "expf2"       "expnf1"      "expnf2"
## [11] "fe_expf2"    "final"       "industries"  "inter"       "lm2_5"
## [16] "lm4_0"       "lm4_1"       "lm4_2"       "lm4_3"       "lm4_4"
## [21] "lm4_5"       "means"       "model_expf2" "out"         "pNIDS"
## [26] "pids"        "plm4_3"      "plm4_3e"     "re_expf2"    "thuso"
## [31] "zbmi1"       "zbmi2"       "zbmi3"       "zbmi4"       "zbmi5"
## [36] "zbmi6"       "zhfa1"       "zhfa2"       "zhfa20"      "zhfa21"
## [41] "zhfa22"      "zhfa23"      "zhfa3"       "zhfa4"       "zhfa5"
## [46] "zhfa6"       "zhfa7"       "zhfa8"       "zhfa9"       "zwfa1"
## [51] "zwfa10"      "zwfa11"      "zwfa2"       "zwfa20"      "zwfa21"
## [56] "zwfa22"      "zwfa23"      "zwfa3"       "zwfa4"       "zwfa5"
## [61] "zwfa6"       "zwfa7"       "zwfa8"       "zwfa9"       "zwfh1"
## [66] "zwfh2"       "zwfh3"       "zwfh4"       "zwfh5"       "zwfh6"
## [71] "zwfh7"       "zwfh8"
```

This data is set up in order to illustrate the benefits of the decompositions. We do this by following the flows of intermediate goods through a fictional GVC and by showing how the readily available gross trade flows differ from the decomposed value added flows. To this end, we construct the elements of the input-output tables such that we have two countries and two industries that

focus on upstream tasks, which means they focus on supplying other industries, and one country and industry that is specialized in downstream tasks, i.e. it serves mainly final demand. In our example the upstream industries are Agriculture and Textiles while the downstream industry is Transport Equipment. Similarly, Argentina and Turkey represent upstream countries with Germany being located downstream within this specific value chain (see table **??**).

The first step of the analytical process is to load the input object and create a decompr class object, which contains the data structures for the decompositions. This step is not needed when using the `decomp()` wrapper function but more on this later.

```
# create the decompr object
decompr_object <- load_tables_vectors( x = inter,
                                       y = final,
                                       k = countries,
                                       i = industries,
                                       o = out        )

# inspect the content of the decompr object
ls(decompr_object)

##  [1] "A"         "Ad"        "Am"        "B"         "Bd"
##  [6] "Bm"        "E"         "ESR"       "Efd"       "Eint"
## [11] "Exp"       "G"         "GN"        "L"         "N"
## [16] "Vc"        "Vhat"      "X"         "Y"         "Yd"
## [21] "Ym"        "bigrownam" "fdc"       "i"         "k"
## [26] "rownam"    "z"         "z01"       "z02"
```

As can be seen above, a decompr class object is in fact a list containing thirty different objects. For example, *Eint* is an object that collects the intermediate goods exports of the industries, while Y refers to the final demand that the industries supply. Depending on the choice of the decomposition, all or some of these objects are used.

3.3 Leontief decomposition

Let us now turn to the algorithms, starting with the Leontief decomposition. We shortly describe the theoretical derivation of the method to expose the internal steps of the decompr package. Afterwards, we turn to the technical implementation and, finally, we describe the output.

Table 3.1: Example Input-Output Table: Leather

		Argentina			Turkey			Germany			Final Demand			Output
		Agriculture	Textile and Leather	Transport Equipment	Agriculture	Textile and Leather	Transport Equipment	Agriculture	Textile and Leather	Transport Equipment	Argentina	Turkey	Germany	
Country	Industry													
Argentina	Agriculture	16.1	5.1	1.8	3.2	4.3	0.4	3.1	2.8	4.9	21.5	6.1	8.4	77.7
Argentina	Textile.and.Leather	2.4	8.0	3.2	0.1	3.2	1.6	1.2	3.9	11.5	16.2	1.9	5.1	58.3
Argentina	Transport.Equipment	0.9	0.5	4.0	0.0	0.1	0.3	0.0	0.4	0.5	11	0.5	0.8	19.0
Turkey	Agriculture	1.1	1.9	0.2	18.0	13.2	6.1	9.0	3.1	8.9	7.5	29.5	14.2	112.7
Turkey	Textile.and.Leather	0.3	2.8	0.1	6.1	28.1	6.3	2.1	2.5	25.6	8.9	24.9	16.9	124.6
Turkey	Transport.Equipment	0.0	0.1	0.3	4.1	3.2	8.9	0.2	0.0	1.8	1.2	18.5	4.9	43.2
Germany	Agriculture	1.2	4.2	0.3	4.1	1.2	0.6	29.0	19.5	17.9	9.2	17.9	51.2	156.3
Germany	Textile.and.Leather	1.3	1.1	0.0	3.2	4.8	2.6	5.1	29.1	24.1	7.9	10.1	38.5	127.8
Germany	Transport.Equipment	2.1	1.4	3.0	4.1	3.1	3.9	11.3	8.1	51.3	25.1	35.2	68.4	217.0

3.3.1 Theoretical derivation

The tools to derive the source decomposition date back to Leontief (1936) who showed that, with a set of simple calculations, national Input-Output tables based on gross terms give the true value added flows between industries. The idea behind this insight is that the production of industry i's output requires inputs of other industries and i's own value added. The latter is the direct contribution of i's output to domestic value added. The former refers to the first round of i's indirect contribution to domestic value added since the input from other industries that i requires for its own production triggers the creation of value added in the supplying industries. As supplying industries usually depend on inputs from other industries, this sets in motion a second round of indirect value added creation in the supplying industries of the suppliers, which is also caused by i's production. This goes on until value added is traced back to the original suppliers and can mathematically be expressed as

$$VB = V + VA + VAA + VAAA + ... = V(I + A + A^2 + A^3 + ...), \quad (3.1)$$

which, as an infinite geometric series with the elements of $A < 1$, simplifies to

$$VB = V(I - A)^{-1}, \quad (3.2)$$

where V is a NxN matrix with the diagonal representing the direct value added contribution of N industries, A is the Input-Output coefficient matrix with dimension NxN, i.e. it gives the direct input flows between industries required for 1$ of output, and $B = (I - A)^{-1}$ is the so called Leontief inverse. VB gives thus a NxN matrix of so called value added multipliers, which denote the amount of value added that the production of an industry's 1$ of output or exports brings about in all other industries. Looking from the perspective of the supplying industries, the matrix gives the value added that they contribute to the using industry's production. If we multiply it with a NxN matrix whose diagonal specifies each industry's total output or exports, we get value added origins as absolute values instead of shares.

The application of the Leontief insight to ICIOs as opposed to national Input-Output tables for our Leontief decomposition is straightforward. V refers now to a vector of direct value added contributions of all industries across the different countries. Its dimension is correspondingly $1xGN$, where G is the number of countries. A is now of dimension $GNxGN$ and gives the industry flows including cross border relationships. Since we are interested in the value added origins of exports we multiply these two matrices with a $GNxGN$ matrix whose diagonal

we fill with each industry's exports, E, such that the basic equation behind the source decomposition is given by $V(I-A)^{-1}E$. [2] In a simple example with two countries (k and l) and industries (i and j) we can zoom in to see the matrices' content:

$$
\begin{aligned}
V(I-A)^{-1}E &= \begin{pmatrix} v_k^i & 0 & 0 & 0 \\ 0 & v_k^j & 0 & 0 \\ 0 & 0 & v_l^i & 0 \\ 0 & 0 & 0 & v_l^j \end{pmatrix} * \begin{pmatrix} b_{kk}^{ii} & b_{kk}^{ij} & b_{kl}^{ii} & b_{kl}^{ij} \\ b_{kk}^{ji} & b_{kk}^{jj} & b_{kl}^{ji} & b_{kl}^{jj} \\ b_{lk}^{ii} & b_{lk}^{ij} & b_{ll}^{ii} & b_{ll}^{ij} \\ b_{lk}^{ji} & b_{lk}^{jj} & b_{ll}^{ji} & b_{ll}^{jj} \end{pmatrix} \\
&\quad * \begin{pmatrix} e_k^i & 0 & 0 & 0 \\ 0 & e_k^j & 0 & 0 \\ 0 & 0 & e_l^i & 0 \\ 0 & 0 & 0 & e_l^j \end{pmatrix} \\
&= \begin{pmatrix} v_k^i b_{kk}^{ii} e_k^i & v_k^i b_{kk}^{ij} e_k^j & v_k^i b_{kl}^{ii} e_l^i & v_k^i b_{kl}^{ij} e_l^j \\ v_k^j b_{kk}^{ji} e_k^i & v_k^j b_{kk}^{jj} e_k^j & v_k^j b_{kl}^{ji} e_l^i & v_k^j b_{kl}^{jj} e_l^j \\ v_l^i b_{lk}^{ii} e_k^i & v_l^i b_{lk}^{ij} e_k^j & v_l^i b_{ll}^{ii} e_l^i & v_l^i b_{ll}^{ij} e_l^j \\ v_l^j b_{lk}^{ji} e_k^i & v_l^j b_{lk}^{jj} e_k^j & v_l^j b_{ll}^{ji} e_l^i & v_l^j b_{ll}^{jj} e_l^j \end{pmatrix} \\
&= \begin{pmatrix} vae_{kk}^{ii} & vae_{kk}^{ij} & vae_{kl}^{ii} & vae_{kl}^{ij} \\ vae_{kk}^{ji} & vae_{kk}^{jj} & vae_{kl}^{ji} & vae_{kl}^{jj} \\ vae_{lk}^{ii} & vae_{lk}^{ij} & vae_{ll}^{ii} & vae_{ll}^{ij} \\ vae_{lk}^{ji} & vae_{lk}^{jj} & vae_{ll}^{ji} & vae_{ll}^{jj} \end{pmatrix}
\end{aligned}
\tag{3.3}
$$

$$
v_c^s = \frac{va_c^s}{y_c^s} = 1 - a_{kc}^{is} - a_{kc}^{js} - a_{lc}^{js} - a_{lc}^{is} \qquad (c \in k,l \quad s \in i,j),
$$

$$
\begin{pmatrix} b_{kk}^{ii} & b_{kk}^{ij} & b_{kl}^{ii} & b_{kl}^{ij} \\ b_{kk}^{ji} & b_{kk}^{jj} & b_{kl}^{ji} & b_{kl}^{jj} \\ b_{lk}^{ii} & b_{lk}^{ij} & b_{ll}^{ii} & b_{ll}^{ij} \\ b_{lk}^{ji} & b_{lk}^{jj} & b_{ll}^{ji} & b_{ll}^{jj} \end{pmatrix} = \begin{pmatrix} 1-a_{kk}^{ii} & -a_{kk}^{ij} & -a_{kl}^{ii} & -a_{kl}^{ij} \\ -a_{kk}^{ji} & 1-a_{kk}^{jj} & -a_{kl}^{ji} & -a_{kl}^{jj} \\ -a_{lk}^{ii} & -a_{lk}^{ij} & 1-a_{ll}^{ii} & -a_{ll}^{ij} \\ -a_{lk}^{ji} & -a_{lk}^{jj} & -a_{ll}^{ji} & 1-a_{ll}^{jj} \end{pmatrix}^{-1},
$$

and

$$
a_{cf}^{su} = \frac{inp_{cf}^{su}}{y_f^u} \qquad (c,f \in k,l \quad s,u \in i,j),
$$

where v_s^c gives the share of industry s's value added, va_c^s, in output, y_s^c, and e_k^i indicates gross exports. Finally, a_{su}^{cf} denotes the share of inputs, inp_{su}^{cf}, in output. The elements of the $V(I-A)^{-1}E$ or vae matrix are our estimates for the country-industry level value added origins of each country-industry's exports.

[2] When using the leontief_output function, the value added multiplier is instead multiplied with each industry's output.

decompr implements this algorithm into R to automate the process of deriving the matrix. Equipped with it, researchers can calculate standard GVC indicators. Examples include Hummels et al. (2001)'s Vertical Specialisation ratio at the industry-level using the `vertical_specialisation` function, which sums for each country and industry across the value added of all foreign countries and industries, and Johnson and Noguera (2012)'s so-called VAX ratio. Alternatively, the four dimensions of the matrix (source country, source industry, using country, using industry) allow for industry-level gravity-type estimations of value added trade flows.

3.3.2 Implementation

As described, in section §3.2, the first step of our analytical process is to construct a decompr object using the `load_tables_vectors()` function. After this, we can use the `leontief()` function to apply to Leontief decomposition.

```
lt <- leontief( decompr_object )
```

In addition, a wrapper function called `decomp()` is provided which integrates both elements of the workflow into a single function. We recommended that the atomic functions be used for large data sets, however, for small data sets this is an easy way to derive the results immediately. The `decomp()` function requires a method to be specified (see `help('decomp')` for details), if none is provided, the function will default to `leontief()`.

```
lt2 <- decomp( x = inter,
               y = final,
               k = countries,
               i = industries,
               o = out,
               method = "leontief" )
```

Note that the output produced by these two different processes is identical.

3.3.3 Output

We can now analyse the output of the Leontief decomposition, which consists of a $GNxGN$ matrix that gives for each country and industry the value added origins of its exports by country and industry. To this end, we look at the results of the Leontief decomposition for our example data set (table 3.2). In the first column we find the source countries and industries while the first row contains the using countries and industries. The first element, 28.52, thus gives the amount of value

added that the Argentinian Agriculture industry has contributed to the exports of the Argentinian Agriculture industry. Similarly, the last element of this row, 4.12, gives the amount of value added that the Argentinian Agriculture industry has contributed to the exports of the German Transport Equipment industry.

A key advantage of the decomposition becomes clear when we compare the decomposed values with the intermediate trade values of the non-decomposed IO table when multiplied with the exports over output ratio to create comparability (table 3.2). We see for instance that Argentina's Agriculture industry contributes significantly more value added to the German Transport Equipment industry than suggested by the non-decomposed IO table. The reason is that Argentina's Agriculture industry is an important supplier to Turkey's Textile and Leather industry which is in turn an important supplier for the German Transport Equipment industry. The decomposition thus allows us to see how the value added flows along this Global Value Chain.

We can also take look at specific industries. For instance, we find that the non-decomposed values of the Transport Equipment are for many elements larger than the value added elements while the opposite holds for Agriculture. This emphasises the fact that Transport Equipment is a downstream industry that produces mostly final goods. Agriculture on the other hand qualifies as an upstream industry that produces also many intermediate goods so that its value added in other industries is typically large.

Finally let's consider the countries of our specific example. We see that Germany has more instances in which the non-decomposed values are above the value added flows than Argentina and Turkey combined. Along the lines of the industry analysis, this shows that Germany focuses within this GVC on downstream tasks producing mostly final goods that contain value added from countries located more upstream. In our example these are Turkey and Argentina.

3.4 Wang-Wei-Zhu decomposition

The Wang-Wei-Zhu decomposition builds upon the Leontief insight but uses, in addition, further valuable information provided in ICIOs. More specifically, the Leontief decomposition traces the value added back to where it originates but ICIOs also contain data on how the value added is subsequently used. This information is extracted by the Wang-Wei-Zhu decomposition, which thereby allows a much more detailed look at the structures of international production networks and the respective positions of countries and industries within them.

Table 3.2: Non-decomposed Values

	Argentina. Agriculture	Argentina. Textile.and. Leather	Argentina. Transport. Equipment	Turkey. Agriculture	Turkey. Textile.and. Leather	Turkey. Transport. Equipment	Germany. Agriculture	Germany. Textile.and. Leather	Germany. Transport. Equipment
Argentina.Agriculture	6.88	2.49	0.25	1.30	2.04	0.08	0.77	0.68	1.76
Argentina.Textile.and.Leather	1.03	3.91	0.44	0.04	1.52	0.31	0.30	0.95	4.13
Argentina.Transport.Equipment	0.38	0.24	0.55	0.00	0.05	0.06	0.00	0.10	0.18
Turkey.Agriculture	0.47	0.93	0.03	7.33	6.27	1.20	2.23	0.75	3.19
Turkey.Textile.and.Leather	0.13	1.37	0.01	2.48	13.35	1.24	0.52	0.61	9.19
Turkey.Transport.Equipment	0.00	0.05	0.04	1.67	1.52	1.75	0.05	0.00	0.65
Germany.Agriculture	0.51	2.05	0.04	1.67	0.57	0.12	7.18	4.73	6.43
Germany.Textile.and.Leather	0.56	0.54	0.00	1.30	2.28	0.51	1.26	7.06	8.65
Germany.Transport.Equipment	0.90	0.68	0.41	1.67	1.47	0.77	2.80	1.96	18.42

Table 3.3: Leontief Decomposition

	Argentina. Agriculture	Argentina. Textile.and. Leather	Argentina. Transport. Equipment	Turkey. Agriculture	Turkey. Textile.and. Leather	Turkey. Transport. Equipment	Germany. Agriculture	Germany. Textile.and. Leather	Germany. Transport. Equipment
Argentina.Agriculture	28.52	2.79	0.36	1.81	3.12	0.36	1.24	1.30	4.12
Argentina.Textile.and.Leather	1.06	19.12	0.42	0.48	1.83	0.43	0.59	1.15	4.75
Argentina.Transport.Equipment	0.21	0.14	1.06	0.03	0.08	0.04	0.02	0.07	0.19
Turkey.Agriculture	0.72	1.34	0.12	34.93	7.00	1.48	2.55	1.52	6.18
Turkey.Textile.and.Leather	0.41	1.39	0.12	2.69	40.17	1.32	1.11	1.15	9.51
Turkey.Transport.Equipment	0.03	0.09	0.03	0.81	0.91	3.16	0.12	0.07	0.65
Germany.Agriculture	0.93	2.25	0.16	2.31	2.06	0.51	29.88	5.25	9.60
Germany.Textile.and.Leather	0.65	0.73	0.08	1.54	2.55	0.63	1.46	18.96	8.16
Germany.Transport.Equipment	0.67	0.65	0.26	1.29	1.49	0.57	1.73	1.51	34.74

3.4.1 Theoretical derivation

The derivation of the Wang-Wei-Zhu decomposition is significantly more technical than the source decomposition since it splits gross exports up more finely. This is why we present here only the final equation for a two country one industry model (equation 22 in WWZ) and refer the interested reader to the original paper by Wang et al. (2014). The key idea is to use the Leontief insight and extend it using additional information from ICIOs on the final usage and destination of the exports (e.g. re-imported vs. absorbed abroad).

$$
\begin{aligned}
E^{kl} = & \left(V^k B^{kk}\right)^T * F^{kl} + \left(V^k L^{kk}\right)^T * \left(A^{kl} B^{ll} F^{ll}\right) \\
& + \left(V^k L^{kk}\right)^T * (A^{kl} \sum_{t \neq k,l}^{G} B^{lt} F^{tt}) + \left(V^k L^{kk}\right)^T * (A^{kl} B^{ll} \sum_{t \neq k,l}^{G} F^{lt}) \\
& + \left(V^k L^{kk}\right)^T * (A^{kl} \sum_{t \neq k}^{G} \sum_{l,u \neq k,t}^{G} B^{lt} F^{tu}) + \left(V^k L^{kk}\right)^T * \left(A^{kl} B^{ll} F^{lk}\right) \\
& + \left(V^k L^{kk}\right)^T * (A^{kl} \sum_{t \neq k,l}^{G} B^{lt} F^{tk}) + \left(V^k L^{kk}\right)^T * \left(A^{kl} B^{lk} F^{kk}\right) \\
& + \left(V^k L^{kk}\right)^T * (A^{kl} \sum_{t \neq k}^{G} B^{lk} F^{kt}) + \left(V^k B^{kk} - V^k L^{kk}\right)^T * \left(A^{kl} X^l\right) \\
& + \left(V^l B^{lk}\right)^T * F^{kl} + \left(V^l B^{lk}\right)^T * \left(A^{kl} L^{ll} F^{ll}\right) + \left(V^l B^{lk}\right)^T \\
& * \left(A^{kl} L^{ll} E^{l*}\right) + (\sum_{t \neq k,l}^{G} V^t B^{tk})^T * F^{kl} + (\sum_{t \neq k,l}^{G} V^t B^{tk})^T \\
& * \left(A^{kl} L^{ll} F^{ll}\right) + (\sum_{t \neq k,l}^{G} V^t B^{tk})^T * \left(A^{kl} L^{ll} E^{l*}\right),
\end{aligned}
\tag{3.4}
$$

where F^{kl} is the final demand in l for goods of k, L^{ll} refers to the national Leontief inverse as opposed to the Inter-Country inverse B, and T indicates a matrix transpose operation. As can be seen from equation (3.4), the Wang-Wei-Zhu decomposition splits gross exports into 16 terms with three main categories given by domestic value added in exports ($DViX\_B$), foreign value added in exports (FVA), and purely double counted terms (PDC). The main categories are further divided according to their final destination so that the final decomposition is given by:

- Domestic value added absorbed abroad (VAX_G, T1-5)
  - Domestic value added in final exports (DVA_FIN, T1)
  - Domestic value added in intermediate exports

        * Domestic value added in intermediate exports absorbed by direct importers (*DVA_INT*, T2)
        * Domestic value added in intermediate exports re-exported to third countries (*DVA_INTrex*, T3-5)
            · Domestic value added in intermediate exports re-exported to third countries as intermediate goods to produce domestic final goods (*DVA_INTrexI1*, T3)
            · Domestic value added in intermediate exports re-exported to third countries as final goods (*DVA_INTrexF*, T4)
            · Domestic value added in intermediate exports re-exported to third countries as intermediate goods to produce exports (*DVA_INTrexI2*, T5)

- Domestic value added returning home (*RDV_B*, T6-8)
    - Domestic value added returning home as final goods (*RDV_FIN*, T6)
    - Domestic value added returning home as final goods through third countries (*RDV_FIN2*, T7)
    - Domestic value added returning home as intermediate goods (*RDV_INT*, T8)

- Foreign value added (*FVA*, T11-12/14-15)
    - Foreign value added in final good exports (*FVA_FIN*, T11/14)
        * Foreign value added in final good exports sourced from direct importer (*MVA_FIN*, T11)
        * Foreign value added in final good exports sourced from other countries (*OVA_FIN*, T14)
    - Foreign value added in intermediate good exports (*FVA_INT*, T12/15)
        * Foreign value added in intermediate good exports sourced from direct importer (*MVA_INT*, T12)
        * Foreign value added in intermediate good exports sourced from other countries(*OVA_INT*, T15)

- Pure double counting (*PDC*, T9-10/13/16)
    - Pure double counting from domestic source (*DDC*, T9-10)
        * Due to final goods exports production (*DDF*, T9)
        * Due to intermediate goods exports production (*DDI*, T10)

  - Pure double counting from foreign source (*FDC*, T13/16)
    * Due to direct importer exports production (*FDF*, T13)
    * Due to other countries' exports production (*FDI*, T16)

The higher resolution of the WWZ decomposition comes at the cost of a lower dimension (source country, using country, using industry) since the current, highly aggregated, ICIOs render a four-dimensional decomposition unfeasible. This means that the two methods are complementary and imply a trade-off between detail and disaggregation.

3.4.2 Implementation

As with the `leontief()` function, the `wwz()` function also takes a decompr class object as its input, the procedure for this is described in section §3.2. After having created this decompr object, we can apply the Wang-Wei-Zhu decomposition using the `wwz()` function.

```
w <- wwz(decompr_object)
```

Furthermore, it is also possible to derive the results of the Wang-Wei-Zhu decomposition directly, using the `decomp()` function.

```
w2 <-  decomp( x = inter,
               y = final,
               k = countries,
               i = industries,
               o = out,
               method = "wwz" )
```

Both these processes will yield the same results.

3.4.3 Output

The output when using the WWZ algorithm is a matrix with dimensions $GNGx19$, whereby 19 consists of the 16 objects the WWZ algorithm decomposes exports into, plus three checksums. GNG represents source country, source industry and using country whereas these terms are slightly ambiguous here due to the complex nature of the decomposition. More specifically, the using country can also be the origin of the foreign value added in the exports of the source country to the using country (see for example T11 and T12). Therefore we use the terms exporter, exporting industry, and direct importer instead. This becomes much clearer when we take a look at specific examples.

table **??** shows the results for the example data. The first column lists exporter, exporting industry, and direct importer. Note that the value added is domestic but not necessarily created in the exporting industry. When exporter and importer are identical, the values are zero since there are no exports. The first row lists the 16 components of bilateral exports at the sector level and three checksums.

The first eight components relate to domestic value added of the exporting country contained in the sectoral exports of the exporting industry to the direct importer. For instance, the first non-zero element in table **??** refers to DVA_FIN, or domestic value added in final good exports. It shows that there are 5.47 units of domestic value added in the exports of final goods from Argentina's Agriculture industry to Turkey. In the same row the third term, *DVA_INTrexI1*, is slightly more complicated. As mentioned above, it gives the amount of domestic Value added in intermediate exports re-exported to third countries as intermediate goods to produce domestic final goods. In our example this means that there are 1.14 units of domestic value added in the intermediate exports of Argentina's Agriculture industry to Turkey, that are re-exported by Turkey as intermediates to a third country which produces final goods with it. Terms six to eight concern domestic value added that eventually returns home. *RDV_FIN2* reveals for example that there are 0.35 units of domestic value added in the intermediate exports of Argentina's Agriculture industry to Turkey, that Turkey re-exports as intermediates to Argentina for the latter's final good production.

The following four terms apply to foreign value added in exports and separate on the one hand between the origin of the foreign value added (MVA vs OVA) and on the other hand between the type of export (intermediate vs final good). MVA describes hereby foreign value added sourced by the exporting country from the direct importer. From the perspective of the latter, these terms are thus part of the RDV (value added returning home) share. OVA in contrast sums over the foreign value added sourced from all other countries. Going back to the example, this means that there are 0.21 units of Turkish value added in the final goods exports of Argentina's Agriculture industry to Turkey.

Terms 13 to 16 collect the double counting of gross trade statistics that occurs when goods cross borders multiple times. DDC captures double counting due to domestic value added, which is further classified according to the type of the ultimate export (final vs intermediate good). MDC and ODC, on the other hand, capture double counting due to foreign value added from either the direct importer or other countries. For the Argentina-Turkey case this implies, for instance, that there are 0.18 units of value added in the intermediate exports of Turkey to Argentina which are re-exported by Argentina's Agriculture industry

to Turkey as intermediates and then again re-exported. Since they would be part of MVA twice, they are now counted once as double-counted term.

Finally, the three checksums give total exports, total final goods exports, and total intermediate exports. The difference between the first and the latter two should be zero.

One interesting application of this decomposition for trade and development uses changes over time in FVA_FIN and FVA_INT. When low-wage developing countries enter GVCs, they tend to specialize mainly in assembly but try to gradually move up within the value chain. To illustrate this, we can reuse the example of the iPhone. Most of the value added in the device stems from US design and Japanese technology but it is ultimately assembled in China. This means for China that when it enters this GVC, its FVA_FIN starts to increase since it imports a lot of foreign value added, assembles it, and exports a final good: the iPhone. However, assembly itself does not contain a lot of value added so that the benefit of China initially is small. When its technology improves due to the interaction with Japan and the USA, it might be able to produce actual parts of the phone, which contain more value added. Eventually it might even be able to outsource assembly to a cheaper country. This would imply that it seizes a larger share of the value added, something commonly referred to as upgrading or moving up within the value chain. In terms of the WWZ decomposition, we would then observe that China's FVA_FIN first goes up and then starts to decline with a simultaneous increase in FVA_INT.

3.5 Conclusion

GVCs describe the increasingly international organization of production structures. As more and more regional trade agreements come into force, which drive down trade costs and harmonize product standards, it becomes more and more attractive for firms to outsource certain tasks of their production lines. Research on international trade analysing this development evolves quickly and reveals important implications of GVCs for economic growth and competitiveness. decompr aims at facilitating this reasearch by simplifying the calculation of standard GVC indicators. The purpose is to accelerate the research and, especially, to make it accesible to a wider audience.

We have designed the package using a modular structure, with an additional user interface function for increased ease of use. The modular structure enables users to break the computationally intensive analysis process down into several steps. Furthermore, the modular structure enables users and other developers to build on top of basic data structures which are created by the `load_tables_vectors` function when implementing other decompositions or

analyses. However, a wrapper function (`decomp()`) is also provided, which combines the use of the atomic functions into one. All of this should allow users of the package to adapt the package to their specific needs as the GVC research progresses.

Since GVCs constitute a fairly new field of research, there are many ways forward for its analysis. The next central step is to examine both in theory and empirically how GVC participation affects real economic activity. More specifically, it is very relevant to look at how, for instance, employment and economic growth react when countries join GVCs and what the factors are that determine a successful relationship. From the standpoint of developing and emerging countries a very interesting question is if GVCs simplify industrialization and the formation of comparative advantage while high-income countries might look for an additional push for their stagnating post-crisis economies. We hope that decompr can play a part in this field and promote it.

Colophon

This paper was written in a combination of R R Core Team (2016) and LaTeX (Lamport and LaTEX, 1986), specifically LuaTeX Hagen 2005, with biblatex and biber (Lehman, 2006) for citations, using Sweave syntax (Leisch, 2003) and compiled using knitr (Xie, 2013) and Pandoc (MacFarlane, 2012).

Bibliography

Hagen, Hans

2005 "LuaTEX: Howling to the moon", *COMMUNICATIONS OF THE TEX USERS GROUP TUGBOAT EDITOR BARBARA BEETON PROCEEDINGS EDITOR KARL BERRY*, p. 152.

Hummels, David, Jun Ishii and Kei-Mu Yi

2001 "The nature and growth of vertical specialization in world trade", *Journal of International Economics*, 54, 1 (June 2001), pp. 75–96, `http://ideas.repec.org/a/eee/inecon/v54y2001i1p75-96.html`.

Johnson, Robert C. and Guillermo Noguera

2012 "Accounting for intermediates: Production sharing and trade in value added", *Journal of International Economics*, 86, 2, pp. 224–236.

Lamport, Leslie and A LaTEX

1986 *Document Preparation System*, Addison-Wesley Reading, MA.

Lehman, Philipp et al.

2006 *The biblatex package.*

Leisch, Friedrich

2003 "Sweave and Beyond: Computations on Text Documents", in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*, ed. by Kurt Hornik, Friedrich Leisch and Achim Zeileis, ISSN 1609-395X, `http://www.R-project.org/conferences/DSC-2003/Proceedings/`.

Leontief, Wassily

1936 "Quantitative Input and Output Relations in the Economic System of the United States", *Review of Economics and Statistics*, 18, 3, pp. 105–125.

MacFarlane, John

2012 *Pandoc: a universal document converter*.

R Core Team

2016 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`.

Timmer, Marcel, AA Erumban, R Gouma, B Los, U Temurshoev, GJ de Vries and I Arto

2012 "The world input-output database (WIOD): contents, sources and methods", *WIOD Background document available at www. wiod. org*, 40.

Wang, Zhi, Shang-Jin Wei and Kunfu Zhu

2014 "Quantifying international production sharing at the bilateral and sector levels".

Wickham, Hadley

2014 *Advanced R*, The R Series, Taylor & Francis Group, Boca Raton, USA, ISBN: 978-1-4665-8696-3, `http://adv-r.had.co.nz/` (visited on 04/02/2014).

Xie, Yihui

2013 *Dynamic Documents with R and knitr*, Chapman and Hall, ISBN: 978-1482203530.

Chapter 4

Global Value Chains in LICs

with Victor Kummritz

Abstract

Global Value Chains are ...

4.1 Introduction

The emergence of Global Value Chains (GVCs) offers a new path to industrialisation for developing countries. As Baldwin (2012) phrases it, internationally fragmented production allows developing countries to join existing supply chains instead of building them. This brings about many potential advantages for these countries. Connecting with firms from advanced nations allows developing nations, for instance, to benefit from their sophisticated technologies and know-how. In addition, relying on an existing production network frees them from constraints imposed by economies of scale and the increased specialisation that GVCs imply limits the negative impact of unproductive parts of the domestic supply chain. After all, when competition moves from goods to tasks, comparative advantage becomes much finer and does not require a broad range of productive stages domestically. Conditional evidence for such a positive impact of GVC participation in low- and middle-income countries is presented in Kummritz (2016) and UNCTAD (2013).

Empirically, the considerable expansion of GVCs has been documented in several recent studies. For instance, Hummels, Ishii et al. (2001); Hummels, Rapoport et al. (1998) show in two early seminal contributions that GVCs are responsible for a major share of the total growth in world trade from 1970 to

1990. Amongst others, Johnson and Noguera (2012a) and Baldwin and Lopez-Gonzalez (2013) find that this growth in GVC trade has even accelerated in the recent two decades. Furthermore, this work has not only revealed a rapid rise in production fragmentation across borders but it has also re-evaluated important indicators of trade, such as bilateral trade imbalances and revealed comparative advantage showing that calculating GVC indicators is central to a better understanding of countries' trade patters and competitiveness.

A central step towards a more in-depth analysis of GVCs has been laid by Koopman et al. (2014) and Wang et al. (2013) who show that it is necessary to go beyond deriving origins of value added to examine production sharing comprehensively. They split goods into different categories and calculate metrics of how often these goods cross borders. This enables them to derive measures of GVC length but also allows to investigate how individual countries are integrated into GVCs. For instance, they show that a considerable part of US value added exports eventually returns home in the form of final goods which is indicative of the US being specialised in upstream production stages.[1]

However, these contributions typically have one of two shortcomings. Firstly, most evidence is based on data from high-income countries. The reason is that reliable time-series of both national and international input-output tables have only been available for this particular subset of countries. In addition, the evidence is regularly based on a small sample of GVC indicators that hide valuable information stemming from more decomposed and disaggregated indicators.

In this paper we address these issues by applying the novel and more detailed gross export decomposition developed by Wang et al. (2013) and Koopman et al. (2014) to a new set of Inter-Country Input-Output tables (ICIOs) with extensive country coverage provided by the OECD. The new ICIOs allow us to get a better understanding of the GVC activities of low- and middle-income countries while the new decomposition allows us to zoom in more closely at these activities revealing information not available from standard GVC indicators.

Our analysis confirms the expansion of GVCs in recent years and presents evidence that GVCs have become longer over time. We also find that these developments are increasingly driven by low- and middle-income countries while the integration of high-income countries has begun to even out at a high level. In addition, we find that high-income countries typically are the starting and end points of GVCs in that they provide upstream inputs and then serve eventually again as demand markets for the final products. Low- and middle-income countries, on the other hand, are more specialised in downstream activities such as assembly and export typically less domestic value added. However, we ob-

[1] See Amador and Cabral (forthcoming) for a comprehensive review of the literature on GVCs and outsourcing.

serve that developing economies have begun to move upstream along the value chain and out of pure assembly occupying a wider set stages. This should allow them to generate greater gains from GVC participation.

The paper is organised as follows. Section 4.2 briefly reviews the decomposition proposed by Wang et al. (2013, WWZ henceforth) and outlines the new ICIOs provided by the OECD. Section 4.3 discusses results using standard indicators and measures calculated with the new data while section 4.4 discusses the results for the novel indicators. Section 4.5 concludes.

4.2 New data and new indicators[2]

GVC analysis relies typically on Inter-Country Input-Output tables (ICIOs). ICIOs are matrices that give supply and demand relationships between industries within and across countries. For instance, ICIOs state the amount of inputs of the Indian steel industry in the output of the US car industry. However, for a correct examination of GVCs it is necessary to go a step further from the ICIOs, by deriving the true value added origins of the US car output. If, for example, India depends on inputs from the US steel industry to supply the US car industry, then ICIOs overstate the actual contribution of India. The extension of the basic Leontief (1936) insight by Hummels, Ishii et al. (2001) shows how the information in ICIOs can be decomposed to estimate such value added flows.

The idea is that the production of industry i of country k creates value added in industry i itself, a direct contribution, but also in industries j from k or other countries l that supply i with inputs, an indirect contribution. Since these industries themselves rely on inputs, i's production sets several rounds of indirect value added creation in motion that can mathematically be expressed as:

$$VB = V + VA + VAA + VAAA + ... = V(I + A + A^2 + A^3 + ...), \qquad (4.1)$$

which, as an infinite geometric series with the elements of $A < 1$, simplifies to

$$VB = V(I - A)^{-1}, \qquad (4.2)$$

where V is a matrix with the diagonal representing the direct value added contribution of each industry, A is the Input-Output coefficient matrix, which means it gives the direct input flows between industries required for 1$ of output, and $B = (I-A)^{-1}$ is the so called Leontief inverse. VB thus gives so called value added multipliers, which denote the amount of value added that the production

[2]The following section draws heavily from Wang et al. (2013), Kummritz (2016), and Quast and Kummritz (2015).

of an industry's \$1 of output or exports brings about in all other industries. If we post-multiply VB with exports, we get a matrix, VAE, with the elements being the value added origins of each industry's exports, vae_{ikjl}.

This basic decomposition has been widely used in GVC analysis since it allows the calculation of two informative GVC participation measures. Firstly, a backward linkage indicator that is given by the import content of exports, $i2e$, (Hummels, Ishii et al. (2001)'s Vertical Specialisation) and calculated as follows:

$$i2e_{ik} = \frac{\sum_l \sum_j vae_{jlik}}{exports_{ik}}, \tag{4.3}$$

Secondly, a forward linkage indicator - *e2r* (domestic content in foreign (re-)exports) - which is given by:

$$e2r_{ik} = \frac{\sum_l \sum_j vae_{ikjl}}{exports_{ik}}, \tag{4.4}$$

where $l \neq k$.

These indicators can tell us how much a country is integrated into GVCs and if it acts mainly as a supplier or a user of foreign value added. However, the Leontief decomposition is only informative for the origin and destination of value added while ICIOs also contain info on the type of good that is being traded and and how often an intermediate crosses borders. The WWZ decomposition extends the Leontief decomposition in this direction and thereby extracts more insights from ICIOs.

4.2.1 Wang-Wei-Zhu decomposition

Since the derivation itself is not the focus of this paper, here we only present the final result for a G-country N-industry model (equation 37 in WWZ) and refer the interested reader to the original paper. WWZ use the Leontief decomposition and extend it using additional information from ICIOs on the final usage and destination of the exports (e.g. re-imported vs. absorbed abroad). This splits the exports, E, of industry l in country k into sixteen different parts broadly differentiated into the four broad categories domestic value added absorbed abroad, domestic value added returning home, foreign value added, and

purely double counted terms:

$$
\begin{aligned}
E^{kl} = & \left(V^k B^{kk}\right)^T * F^{kl} + \left(V^k L^{kk}\right)^T * \left(A^{kl} B^{ll} F^{ll}\right) \\
& + \left(V^k L^{kk}\right)^T * (A^{kl} \sum_{t \neq k,l}^{G} B^{lt} F^{tt}) + \left(V^k L^{kk}\right)^T * (A^{kl} B^{ll} \sum_{t \neq k,l}^{G} F^{lt}) \\
& + \left(V^k L^{kk}\right)^T * (A^{kl} \sum_{t \neq k}^{G} \sum_{l,u \neq k,t}^{G} B^{lt} F^{tu}) + \left(V^k L^{kk}\right)^T * \left(A^{kl} B^{ll} F^{lk}\right) \\
& + \left(V^k L^{kk}\right)^T * (A^{kl} \sum_{t \neq k,l}^{G} B^{lt} F^{tk}) + \left(V^k L^{kk}\right)^T * \left(A^{kl} B^{lk} F^{kk}\right) \\
& + \left(V^k L^{kk}\right)^T * (A^{kl} \sum_{t \neq k}^{G} B^{lk} F^{kt}) + \left(V^k B^{kk} - V^k L^{kk}\right)^T * \left(A^{kl} X^l\right) \\
& + \left(V^l B^{lk}\right)^T * F^{kl} + \left(V^l B^{lk}\right)^T * \left(A^{kl} L^{ll} F^{ll}\right) + \left(V^l B^{lk}\right)^T \\
& * \left(A^{kl} L^{ll} E^{l*}\right) + (\sum_{t \neq k,l}^{G} V^t B^{tk})^T * F^{kl} + (\sum_{t \neq k,l}^{G} V^t B^{tk})^T \\
& * \left(A^{kl} L^{ll} F^{ll}\right) + (\sum_{t \neq k,l}^{G} V^t B^{tk})^T * \left(A^{kl} L^{ll} E^{l*}\right),
\end{aligned}
\tag{4.5}
$$

where F is final demand, and L refers to the domestic Leontief inverse as opposed to the global inverse B. X is output while T indicates a matrix transpose operation.

The four main categories are further divided according to their final destination so that the final decomposition is given by:

- Domestic value added absorbed abroad (VAX_G, T1-5)
  - Domestic value added in final exports (DVA_FIN, T1)
  - Domestic value added in intermediate exports (DVA_INT, T2-5)
    - Domestic value added in intermediate exports absorbed by direct importers (DVA_INT, T2)
    - Domestic value added in intermediate exports re-exported to third countries (DVA_INTrex, T3-5)
      - Domestic value added in intermediate exports re-exported to third countries as intermediate goods to produce domestic final goods ($DVA\_INTrexI1$, T3)
      - Domestic value added in intermediate exports re-exported to third countries as final goods ($DVA\_INTrexF$, T4)
      - Domestic value added in intermediate exports re-exported to third countries as intermediate goods to produce exports

($DVA\_INTrexI2$, T5)

- Domestic value added returning home (RDV, T6-8)
  - Domestic value added returning home as final goods (RDV_FIN, T6)
  - Domestic value added returning home as final goods through third countries (RDV_FIN2, T7)
  - Domestic value added returning home as intermediate goods (RDV_INT, T8)
- Foreign value added (FVA, T11-12/14-15)
  - Foreign value added in final good exports (FVA_FIN, T11/14)
    - Foreign value added in final good exports sourced from direct importer (MVA_FIN, T11)
    - Foreign value added in final good exports sourced from other countries (OVA_FIN, T14)
  - Foreign value added in intermediate good exports (FVA_INT, T12/15)
    - Foreign value added in intermediate good exports sourced from direct importer (MVA_INT, T12)
    - Foreign value added in intermediate good exports sourced from other countries(OVA_INT, T15)
- Pure double counting (PDC, T9-10/13/16)
  - Pure double counting from domestic source (DDC, T9-10)
    - Due to final goods exports production (DDF, T9)
    - Due to intermediate goods exports production (DDI, T10)
  - Pure double counting from foreign source (FDC, T13/16)
    - Due to direct importer exports production (FDF, T13)
    - Due to other countries' exports production (FDI, T16)

For the analysis, we use dva_fin, fva_fin, rdv, pdc and the two aggregate measures dva_inter combining dva_int and ddc as well as fva_inter combining fva_int and fdc. This collapses the indicator to a intuitive and manageable amount.

The advantage of such a detailed decomposition is that these new indicators can inform us on how countries integrate into GVCs while the basic Leontief decomposition mainly informs us on the intensity of integration. High amounts of foreign value added in final goods exports are, for instance, suggestive of a specialisation in downstream tasks that add little value to a good, such as

assembly. High amounts of domestic value added in intermediate exports, on the other hand, are evidence of a more upstream specialisation in tasks that add a lot of value, such as business services. By tracking these two variables over time we can see which countries have succeeded in moving up the value chain. We will explain the indicators in more detail in combination with the decomposition results to facilitate the understanding.

Finally, it is necessary to point out that the high resolution of the WWZ decomposition does not mean that the Leontief decomposition does not contain valuable information at all. In fact, we exploit the decomposition of exports into source industry and source country by calculating variants of the standard indicators based on different characteristics. In particular, we will assess the integration of low- and middle-income countries into GVCs by computing the amount of value added that they supply for total GVC trade.

4.2.2 OECD ICIOs

We use the new OECD ICIOs as the main data source for the GVC indicators and the industry position indicators. The OECD ICIOs constitute the most recent and most advanced release of Inter-Country Input-Output tables. The new version of the database provides ICIOs covering 61 countries and 34 industries for the years 1995, 2000, 2005, and 2008 to 2011.[3,4] This extensive country coverage is crucial for analysing how GVCs affect countries at different stages of development over time, a feature that has not been possible due to limited data availability in previous databases. The empirical literature discussed above shows that especially the extended coverage of Asia is important. To create ICIOs, the OECD combines national IO tables with international trade data. As OECD countries have a harmonised construction methodology, potential discrepancies between national IO tables should be minor. Furthermore, the advanced harmonisation across countries reduces the use of proportionality assumptions to derive the ratio of imported intermediates in an industry's demand to a minimum. In addition, the OECD has used elaborate techniques to deal with China's processing trade. Due to China's outstanding role in GVCs and processing trade, this implies a significant improvement for the reliability of the database.[5]

[3] Countries and industries are listed in Appendix **??**.
[4] Note that in the analysis 2009 and 2010 are excluded due to the global financial crisis.
[5] See Koopman et al. (2012) for an analysis of China's processing trade.

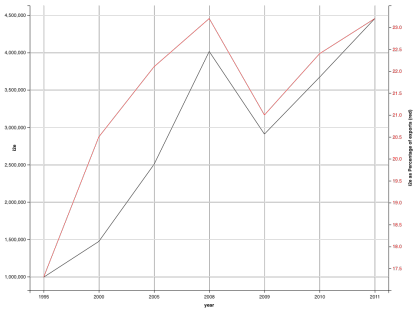


Figure 4.1: The development of GVC integration over time

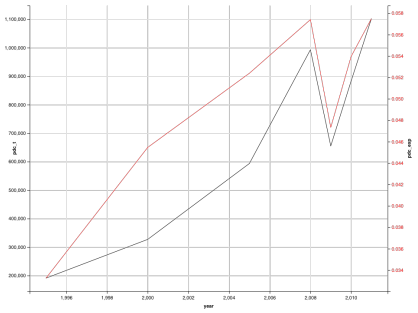


Figure 4.2: The development of double counted trade over time

4.3 What we know: Old facts with new data

In this section we use the extensive OECD ICIO dataset to reasses some stylised facts on GVC integration that are typically based on smaller samples. We start by examining the development of our most basic measure of GVC integration, namely the amount of foreign value added in exports labeled by Baldwin and Lopez-Gonzalez (2013) as $i2e$.[6] It captures backward linkages into value chains and shows the well-known increase in GVC integration from 1995 to 2011. As illustrated in Figure 4.1, the value of $i2e$ has grown by approximately 350% and by 35% as a share of total exports from around 17% to over 23%. Thus, countries rely for their export production increasingly on inputs produced abroad. The numbers are in line with similar findings by Johnson and Noguera (2012b) but their sample ends in 2009. It is then interesting to see that after the slump during the financial crises in 2009, GVCs have quickly recovered and already have started to exceed their pre-crisis levels in 2011.

Another way to examine the expansion of GVCs from 1995 to 2011 is to look at their length instead of their trade volume. WWZ propose to use the amount of double counted trade, *pdc*, as a proxy for GVC length since its value goes up the more back-and-forth trade occurs, which is equivalent to an increase in the number of production stages. They show that its value has increased for 40 selected countries. In Figure 4.2, we observe in our larger sample similarly that *pdc* as a share of total exports has increased over the examined period by 73% and thus more than $i2e$. Therefore, GVCs do not only channel more trade but also have become longer over time.

Turning from the development over time to sectoral differences in GVC integration, Figure 4.3 shows - in line with Johnson and Noguera (2012a) - that the sectors exhibiting the highest degree of international fragmentation in terms of

[6]Note that at the aggregate level forward ($e2r$) and backward ($i2e$) linkages are identical and thus we only look at one of the two measures.

$i2e$ shares are heavy manufactures such as motor vehicles (MTR), other transport equipment (TRQ) and the metal industry (MET) as well as computers and electronics (CEQ and ELQ). In particular, the transport equipment and electronics industry are strongly engaged in GVCs having highly international production networks. For instance, Apple's iPhone contains inputs from 9 to 10 countries while the Boeing 787 production spans more than 5 countries. The sectors can be characterised as being close to final demand and producing complex differentiated goods. These characteristics can thus explain differences in GVC integration.

The bottom 6 industries in terms of $i2e$ shares are primary and services sectors such as agriculture (AGR), mining (MIN), R&D and business services (BZS), or wholesale and retail trade (WRT). These sectors are typically located upstream in the supply chain far from final demand and have high value added to output ratios.

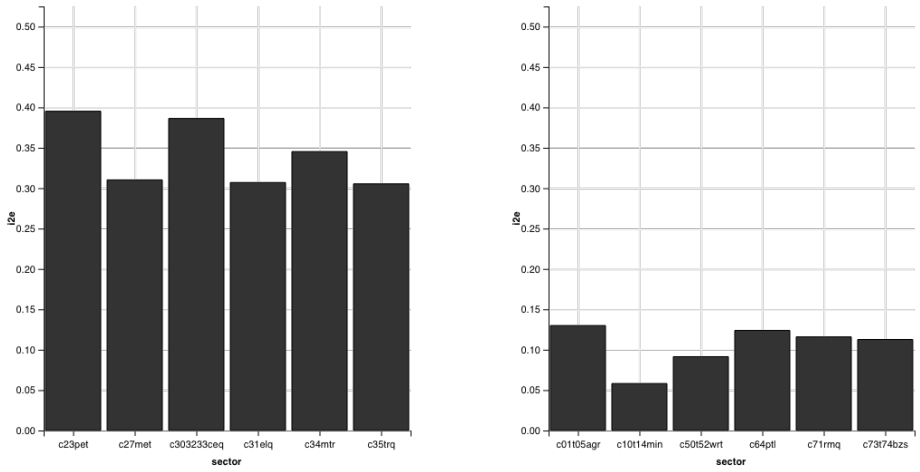


Figure 4.3: Sectoral $i2e$ shares - Top and bottom 6.

Naturally then, things are reversed when we look at the corresponding forward linkage GVC measure, $e2r$. It captures the amount of domestic value added in foreign exports and thus quantifies how important domestic industries are for foreign export production. Here, Figure 4.4 demonstrates that this indicator is dominated by the same upstream industries that are at the bottom of the $i2e$ ranking such as mining or business and telecommunication services (PTL). This shows that these industries are also strongly engaged in GVCs but their participation is of a different type. They primarily supply important inputs, but they do not serve final demand.

The high $e2r$ values of the services sector, also suggest the servicification of manufacturing as described by Baldwin, Forslid et al. (2015). This means that

an increasing share of manufacturing gross exports is actually value added generated in services sectors and then embedded in the intermediate goods exports of manufacturers. This importance of services sectors to exports cannot be seen from standard gross trade statistics and thus constitutes a major advantage of trade in value added measures.

It is also indicative of a growing internationalisation of services. More and more, services are being offshored and sourced from abroad. In that respect, it is also interesting to note that despite the low absolute $i2e$ shares, it is in services where much of the growth in $i2e$ has taken place. Five out of the six sectors with the highest growth in $i2e$ shares are services sectors.

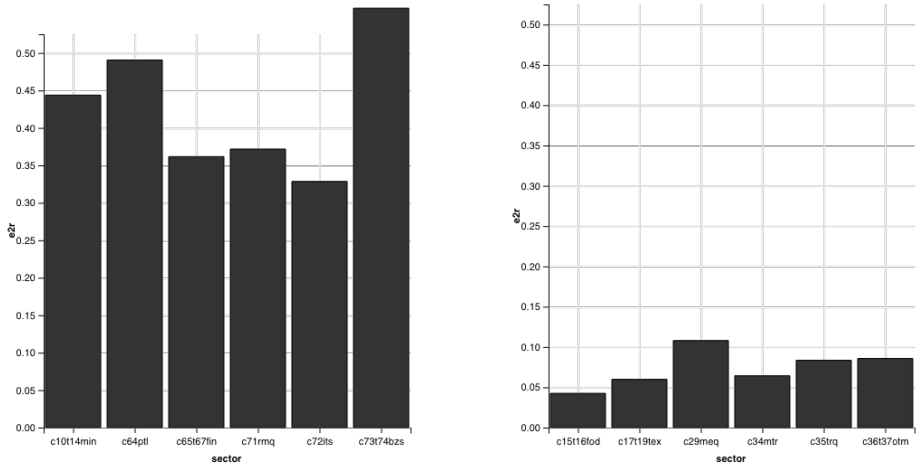


Figure 4.4: Sectoral $e2r$ shares - Top and bottom 6.

Finally, when we turn to differences in GVC integration by country, we can confirm the findings by Baldwin and Lopez-Gonzalez (2013), Figure 4.5 shows that small countries close to the major GVC hubs in Asia, Europe, and North America have the highest average $i2e$ shares. Examples include Malaysia and Slovakia. Countries specialised in the primary sector or assembly on the other hand have very low values. Correspondingly, Latin American countries with their focus on agriculture and mining have very weak backward linkages into GVCs. However, the development over time shows that some of the countries with the relatively low GVC integration have begun to catch up. For instance, Argentina, India and Turkey are in the top 6 when it comes to the growth of $i2e$ shares from 1995 to 2011.

Driven by the sectoral statistics, we then find again that for $e2r$ the picture is reversed with raw material exporters on top. If we abstract from these countries we find technologically advanced countries such as Switzerland and the main GVC hubs Japan, USA, and Germany to exhibit strong forward linkages into

GVCs. In particular low and middle-income countries without raw materials such as Cambodia, Mexico, or Turkey in contrast have very weak linkages and have not been able to strengthen them significantly between 1995 and 2011.[7]

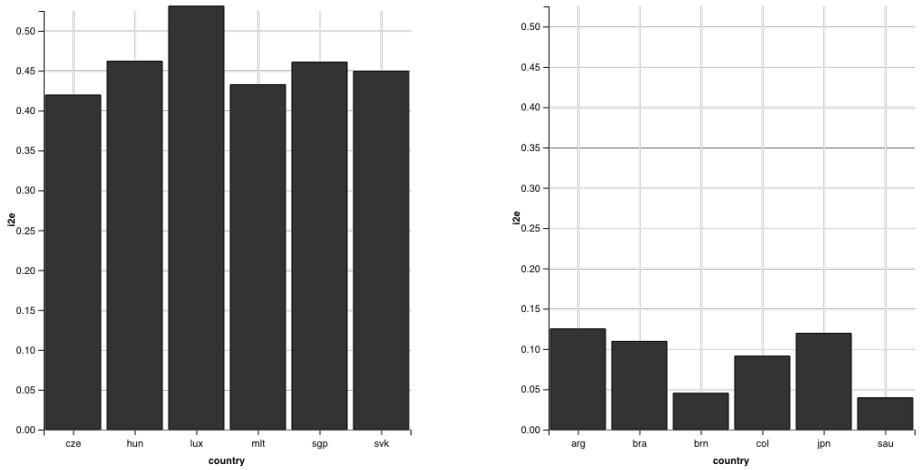


Figure 4.5: Countries' $i2e$ shares - Top and bottom 6.

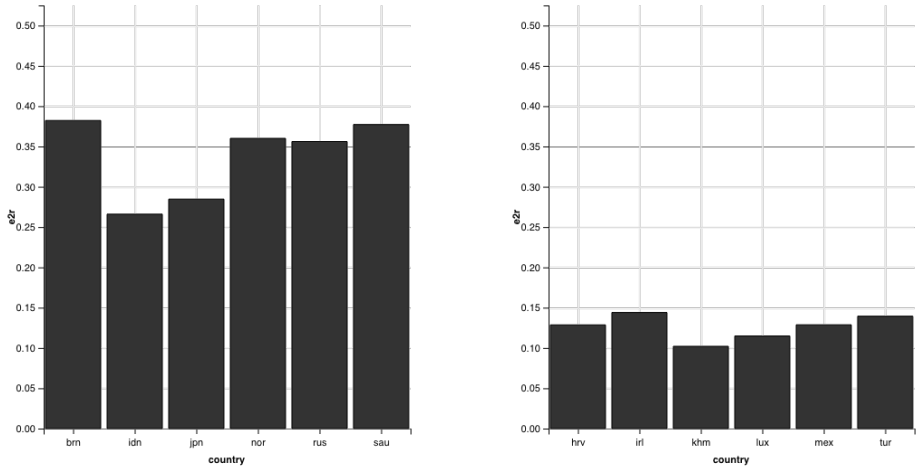


Figure 4.6: Countries' $e2r$ shares - Top and bottom 6.

[7]The full set of results for $i2e$ and $e2r$ by country, and sector can be found in Appendix **??**. Since the results of WWZ decomposition are much more detailed, these results are not presented here are only available from the authors upon request.

4.4 The role of developing economies: New trends and patterns in GVCs

The central advantage of our approach is that we have new indicators for a new set of countries. This means that other than confirming previous findings with a more representative sample, we can also provide several new insights. In particular, the OECD ICIO database extends the available list of countries in ICIOs by the following 21 regions: Argentina, Brunei Darussalam, Cambodia, Chile, Colombia, Costa Rica, Croatia, Hong Kong, Iceland, Israel, Malaysia, Norway, New Zealand, Philippines, Saudi Arabia, Singapore, Thailand, Tunisia, Vietnam, South Africa, and Switzerland. This means that in particular the coverage of low and middle income countries has increased considerably which allows us to analyse the GVC integration of developing economies in a more detailed fashion.

4.4.1 General trends in the GVC participation of developing economies

Regarding the integration of low- and middle-income countries, Johnson and Noguera (2012a) have observed that per capita income is only a weak predictor for GVC integration due to the heterogeneity of economies in terms of size, industrial structure and location. In Table 4.1 we see that the average integration measured by either $i2e$ or $e2r$ does not vary strongly between income groups defined by the World Bank classification at the beginning of the sample period in 1995.[8] High-income economies have slightly stronger forward linkages but lower backward linkages which implies that their exports contain more domestic value added. Developing economies thus have to chance to try to upgrade their GVC integration, by increasing domestic content in exports.

Table 4.1: GVC integration by income

Country group	*i2e*		*e2r*	
	Average	Δ 95-11	Average	Δ 95-11
Low/Lower middle	23.46%	48.22%	20.35%	38.58%
High	22.64%	41.84%	21.85%	29.50%

Data is averaged across countries, sectors and years. Δ 95-11 refers to the growth of the i2e and e2r values from 1995 to 2011.

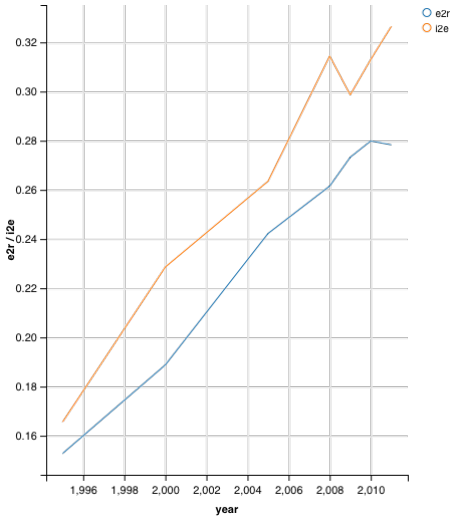
Looking at the development over time, it is striking that the rise of GVC

[8]Note that in this section indicators are based only on manufacturing and services sectors to avoid spurious results stemming from primary sectors that are for technological reasons less integrated into GVCs.

integration is increasingly driven by developing countries. The growth of both $i2e$ or $e2r$ has been much more pronounced in these economies as can be seen in Table 4.1. In relative terms this means that the $i2e$ share of countries classified as low- or lower middle-income in total $i2e$ has increased from 9% in 1995 to 24% in 2011. Similarly, the $e2r$ share has increased from 9% to 23%.

Moreover, low- and lower middle-income countries do not only sell and source more from GVCs but they are also increasingly on the other side of the transaction. Figure 4.7 shows that the share of $i2e$ sourced from low- and lower middle-income countries has risen from 17% to 33% and the share of $e2r$ re-exported from them has expanded from 15% to 28%. Thus, developing countries have a large stake in GVCs and have moved from the periphery into the centre of these production networks.[9]

Figure 4.7: Share of value added sourced from (i2e) or sold to (e2r) low- and lower-middle income economies for export production.



We now zoom in and analyse the GVC participation of developing economies more closely with the help of the WWZ decomposition. As described in section 4.2.1, WWZ show how the structure and changes in the structure of domestic and foreign content in exports inform us about a country's movement along the value chain. In particular, $i2e$ consists of foreign value added in final goods exports (fva_fin), intermediate goods exports (fva_int), and double counting (fdc). Table 4.4.1 shows that on average low- and lower middle-income countries have a higher share of fva_fin in $i2e$ (42%) than high-income economies (39%). This is in line with a trend of specialisation of developing economies in

[9] We will see that GVC integration nevertheless differs significantly among developing countries.

downstream assembly tasks.

Table 4.2: WWZ decomposition results by income

Country group	*fva_fin*	*fva_inter*	*dva_fin*	*dva_inter*	*rdv*
Low/Lower middle	42.07%	57.93%	44.09%	54.73%	1.18%
High	39.38%	60.62%	40.73%	56.85%	2.42%

Data is averaged across countries, sectors and years. Δ 95-11 refers to the growth of the i2e and e2r values from 1995 to 2011.

However, a shift from foreign content in final goods to intermediate goods and double counted trade value would be indicative of moving up the value chain. For low- and lower middle-income countries, we indeed find - as shown by Figure 4.8 - that the share of fva_fin in $i2e$ has fallen by about 4%. This gain accrues to the double counting part, which rises by 6%. This means that production has become more fragmented and that developing economies increasingly occupy more upstream stages of the value chain.

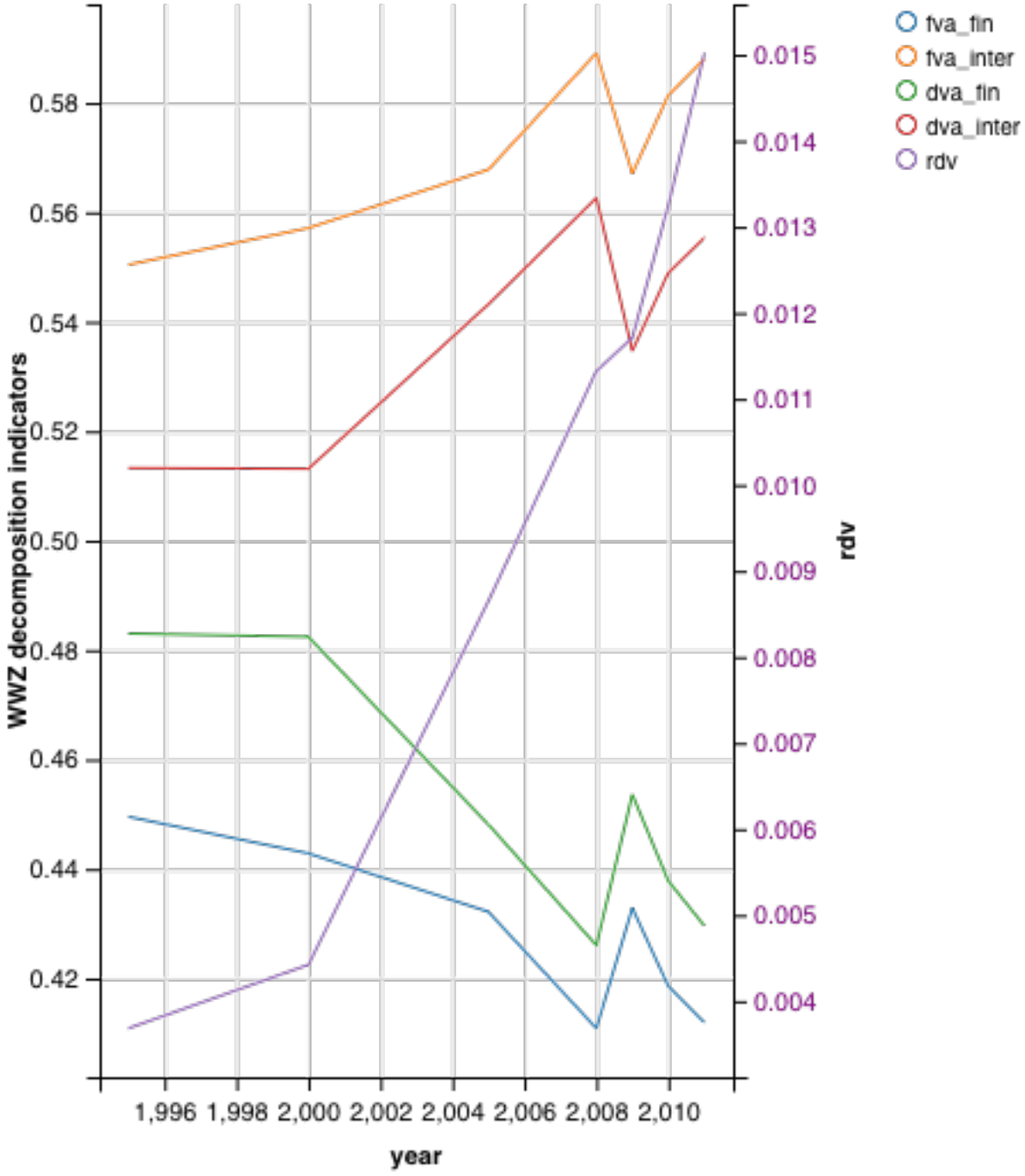
A similar exercise can be done for the domestic value added embodied in exports. The exported domestic value added of high-income countries tends to be dominated by intermediate goods (57%) while low- and lower middle-income countries only achieve a value of 55%. We come to the same conclusion when we look at the share of domestic value added that eventually returns home. Here, the value for high-income countries (2.42%) is more than twice as high than its low- and lower-middle income counterpart (1.18%), which indicates that high-income countries are located upstream in the value chain using developing economies for assembly. However, the data shows as well that developing economies have improved their position over time. The amount of domestic value added returning home has tripled from 1995 to 2011 and the share of final goods has decreased by more than 5%.

Thus, overall we get a clear picture that while developing economies are still positioned relatively more downstream in the value chain, they have succeeded to move up over the past two decades.

4.4.2 Revealing new trends in the participation of developing economies

The trends described in the previous section inform us on the average performance of developing countries but they might hide considerable heterogeneity among these countries, we therefore merge a subset of the newly available countries into the three regions Central and South America (CSA), South East Asia (SEA), and Africa (AFR) and analyse the development of their GVC particip-

Figure 4.8: Development of developing economies' WWZ decomposition indicators over time.



ation country by country. CSA covers Argentina, Chile, Colombia, and Costa Rica; SEA covers Cambodia, Malaysia, The Philippines, Thailand, and Vietnam; while AFR covers South Africa and Tunisia.

South East Asia The SEA economies for which data is newly available are Cambodia, Hong Kong, Malaysia, Philippines, Singapore, Thailand, and Vietnam. Since Singapore and Hong Kong are special cases due to their per capita income and size, we focus on Cambodia, Malaysia, The Philippines, Thailand, and Vietnam.

The two basic indicators of these countries, $i2e$ and $e2r$, presented in Table 4.4.2 show that all five countries are primarily integrated into GVCs through backward linkages but in particular the Philippines have increased their forward linkages over the past two decades considerably. It also stands out that Cambodia and Vietnam have very low $e2r$ values suggesting a strong specialisation in low value added tasks located downstream in the chain. However, in order to obtain more detailed information on how these countries engage in GVCs we need more disaggregated indicators.

The WWZ decomposition provides us with the necessary tools. We can see in Table 4.4.2 that according to their high fva_fin values Cambodia and to a lesser extent Vietnam indeed perform mostly downstream tasks with typically low value added whereas Malaysia, Thailand, and the Philippines are positioned higher in the value chain exhibiting much lower fva_fin and dva_fin but higher rdv values. Comparing these results to the analysis by WWZ, we find that the latter set of countries have a similar GVC integration structure to Indonesia but still lag behind more advanced nations such as Korea and Taiwan.

When we look at the change over time from 1995 to 2011, we see that Cambodia has actually moved into assembly with an increase of fva_fin of 35.2%. This stands in stark contrast to the remaining SEA countries which all managed to move up the value chain. In particular, Vietnam is on a good path with the highest decline of fva_fin and might soon catch up with its local competitors regarding its position in GVCs. For Cambodia, on the other hand, this means that GVCs offer a major untapped potential for future growth. If it is able to introduce more GVC-friendly policies, it can leverage its location close to the GVC hubs China and Japan to put it on a successful growth path.

Table 4.3: GVC integration of SEA countries

Country	$i2e$		$e2r$	
	Average	Δ 95-11	Average	Δ 95-11
Cambodia	39.4%	90.7%	8.4%	-11.9%
Malaysia	44.3%	37.1%	13.9%	10.2%
Philippines	29.6%	-20.7%	22.6%	105.0%
Thailand	36.9%	64.3%	13.1%	20.2%
Vietnam	38.3%	66.1%	10.6%	5.3%

Data is averaged across sectors and years. Δ 95-11 refers to growth from 1995 to 2011.

Table 4.4: WWZ decomposition results for SEA countries

Country	fva_fin		fva_inter		dva_fin		dva_inter		rdv	
	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11
Cambodia	68.1%	35.2%	31.9%	-35.7%	64.5%	26.8%	35.5%	-27.5%	0.0%	-29.7%
Malaysia	39.3%	-9.0%	60.7%	6.3%	40.8%	-4.5%	58.9%	3.4%	0.4%	-21.5%
Philippines	35.5%	-21.7%	64.5%	16.0%	38.9%	-19.1%	60.9%	16.0%	0.2%	18.2%
Thailand	41.4%	-12.9%	58.6%	11.3%	47.4%	-14.6%	52.3%	17.7%	0.3%	20.0%
Vietnam	47.1%	-22.6%	52.9%	30.0%	55.0%	-9.0%	44.8%	12.7%	0.1%	103.4%

Data is averaged across sectors and years. fva variables are expressed as % of $i2e$, dva and rdv variables as % of domestic value added in total exports. Δ 95-11 refers to growth from 1995 to 2011.

Central and South America The newly available CSA economies are Argentina, Chile, Colombia, and Costa Rica, in addition to the previously available Mexico and Brazil. What stands out from looking at the standard GVC indicators presented in Table 4.4.2 is that CSA is on average less integrated into GVCs than SEA and other developing regions. In particular, Argentina and Colombia have both very low backward and forward linkages highlighting the role of remoteness and sound policies as drivers of GVC integration. This is also mirrored in the fact that Chile and Costa Rica exhibit much higher GVC participation rates; albeit still below the SEA countries. These countries perform relatively well in several measures capturing a country's policy environment such as the World Bank's Doing Business Indicators or World Governance Indicators and, in the case of Costa Rica, are relatively closer to the North American GVC centre encompassing the USA, Canada, and Mexico.

When focussing on Costa Rica and Chile, we observe in Table 4.4.2 that Chile's GVC integration structure starts to resembles the structure of high income countries. The largest part of the country's integration is through intermediates as shown by the high fva_inter and dva_inter shares (78% and 75% respectively). However, the share of returned domestic value (rdv) is still much

lower than the high-income average of 2.4% and thus indicates that Chile is still in the process of catching up.

Costa Rica possesses the typical GVC integration structure of lower middle-income economies with high fva_fin and dva_fin shares and a very small rdv value of 0.02%. Comparing the country to SEA, its structure resembles most closely the GVC integration of Vietnam. This comparison holds also when we look at Costa Rica's development over time, where we see a rapid expansion of fva_inter, dva_inter, and rdv shares. The country is thus successfully moving up the value chain.

Table 4.5: GVC integration of CSA countries

Country	$i2e$		$e2r$	
	Average	Δ 95-11	Average	Δ 95-11
Argentina	13.4%	154.9%	13.4%	19.4%
Chile	20.0%	44.8%	26.4%	35.4%
Colombia	13.2%	15.2%	17.0%	45.5%
Costa Rica	29.0%	21.1%	16.0%	60.8%

Data is averaged across sectors and years. Δ 95-11 refers to growth from 1995 to 2011.

Table 4.6: WWZ decomposition results for CSA countries

Country	fva_fin		fva_inter		dva_fin		dva_inter		rdv	
	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11
Argentina	51.15%	-6.33%	48.85%	7.95%	51.92%	-5.71%	47.90%	7.01%	0.18%	49.61%
Chile	22.23%	-22.29%	77.77%	8.54%	24.61%	-23.44%	75.25%	10.03%	0.14%	81.25%
Colombia	39.41%	-19.09%	60.59%	15.30%	39.32%	-33.76%	60.55%	32.06%	0.12%	25.56%
Costa Rica	45.99%	-11.17%	54.01%	11.29%	50.97%	-17.88%	49.01%	24.97%	0.02%	43.05%

Data is averaged across sectors and years. fva variables are expressed as % of $i2e$, dva and rdv variables as % of domestic value added in total exports. Δ 95-11 refers to growth from 1995 to 2011.

Africa To conclude, we turn to Africa. GVC data on Africa is scarce and typically it is assumed that integration levels are low. However, the newly available OECD data includes Tunisia and South Africa, two interesting and unique cases. Tunisia and South Africa offer relatively stable political environments and a relatively high degree of industrialisation which makes them two optimal case studies. Unlike many other African they do thus fulfil the basic requirements for GVC integration.

In line with this, Tables 4.4.2 and 4.4.2 show that in fact Tunisia has relatively high integration levels. Its integration pattern is very similar in both intensity, structure, and trend to Costa Rica and Vietnam. This means that

Tunisia is mainly integrated through backward linkages and assembly tasks but is moving up the value chain. This is evidence that especially North Africa with its proximity to the European GVC hub can link into and benefit from GVCs.

South Africa is a different case since it is located far from most production networks and focuses primarily on raw materials. As a result, the country's integration levels are fairly low and more similar to Argentina and Colombia. Nevertheless, it is likely that it has benefitted from the boom in commodities caused by the rise of GVCs and the subsequent boost in global demand.

Table 4.7: GVC integration of AFR countries

Country	$i2e$		$e2r$	
	Average	Δ 95-11	Average	Δ 95-11
South Africa	21.3%	61.4%	19.9%	16.4%
Tunisia	32.1%	35.6%	13.2%	33.1%

Data is averaged across sectors and years. Δ 95-11 refers to growth from 1995 to 2011.

Table 4.8: WWZ decomposition results for AFR countries

Country	fva_fin		fva_inter		dva_fin		dva_inter		rdv	
	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11	Average	Δ 95-11
South Africa	48.76%	-11.76%	51.24%	13.56%	54.43%	-14.60%	45.49%	21.41%	0.08%	7.07%
Tunisia	45.09%	-14.97%	54.91%	15.19%	56.62%	-4.47%	43.10%	5.56%	0.28%	147.59%

Data is averaged across sectors and years. fva variables are expressed as % of $i2e$, dva and rdv variables as % of domestic value added in total exports. Δ 95-11 refers to growth from 1995 to 2011.

4.5 Conclusion

GVCs are a major new factor in international trade. International production networks span across many countries and affect many industries while changing the way trade impacts domestic economies. This development requires new data and new statistics that appropriately capture countries' integration into GVCs. In this paper, we make use both such novelties in terms of data and statistics by applying a novel gross export decomposition methodology to a new expanded dataset.

More precisely, we apply the Wang-Wei-Zhu decomposition based on Wang et al. (2013) and Koopman et al. (2014) to a new set of Inter-Country Input-Output tables built by the OECD. The advantage is twofold. Firstly, the WWZ decomposition allows us to analyse the structure of regions' GVC integration in addition to the intensity measures provided by previous decompositions leading

to deeper insights into GVC integration patterns. Secondly, the new OECD ICIOs cover a more developing economies than previous ICIOs. This allows us to develop a more accurate understanding of how these countries integrate with GVCs.

We find that many ideas based on previous anecdotal evidence can be confirmed by the data. In particular, there is a central difference in the structure of high-income economies' integration into GVCs compared to developing economies when it comes to the position in GVCs. If we set aside primary sectors, high-income economies are typically positioned more upstream in the value chain which can be seen from the concentration of their value added in intermediate goods exports. In addition, they also serve as market of final demand which can be seen from their relatively high share of exported domestic value added returning home eventually for final consumption.

Developing economies, on the other hand, tend to be positioned more downstream, this can be deduced from the concentration of their GVC participation in final goods exports and the fact that their forward linkages and returning domestic value added tend to be relatively low. These two stylised facts suggest that high-income economies use GVCs to outsource low value added downstream production stages and eventually reimport the final goods. However, when looking at the development over time, it appears that many developing economies have succeeded in moving up the value chain and that the general trend points to a more even distribution of value added across the different countries.

Finally, we use the new data to look at selected low- and middle income economies in three different regions, namely South-East Asia, Latin America and the Caribbean, and Africa. South-East Asia has as expected the highest levels of GVC integration while we observe more heterogeneity in Latin America and the Caribbean where especially Chile and Costa Rica perform well. In Africa, we find that Tunisia has developed backward linkages into GVCs, which shows that Northern Africa has the potential to become part of the European GVC network.

Overall, we show that low- and middle-income countries have become an integral part of GVCs and are increasingly becoming the driver of their expansion. In addition, they increasingly succeed in moving into higher value added stages of the production networks. While the exact implications of integration into GVCs are still the subject of much research, it is clear that they offer significant potential for industrialisation and growth and that countries like The Philippines, Costa Rica, or Tunisia are therefore in good positions to benefit from this and can serve as examples for comparable countries.

Bibliography

Amador, Joao and Sonia Cabral

n.d. "GLOBAL VALUE CHAINS: A SURVEY OF DRIVERS AND MEASURES", *Journal of Economic Surveys*, forthcoming.

Baldwin, Richard

2012 *Global supply chains: Why they emerged, why they matter, and where they are going*, CEPR Discussion Papers 9103, C.E.P.R. Discussion Papers, `http://ideas.repec.org/p/cpr/ceprdp/9103.html`.

Baldwin, Richard, Rikard Forslid and Tadashi Ito

2015 *Unveiling the Evolving Sources of Value Added in Exports*, Joint Research Program Series 161, IDE-JETRO.

Baldwin, Richard and Javier Lopez-Gonzalez

2013 *Supply-Chain Trade: A Portrait of Global Patterns and Several Testable Hypotheses*, NBER Working Papers 18957, National Bureau of Economic Research, Inc, `http://ideas.repec.org/p/nbr/nberwo/18957.html`.

Hummels, David, Jun Ishii and Kei-Mu Yi

2001 "The nature and growth of vertical specialization in world trade", *Journal of International Economics*, 54, 1 (June 2001), pp. 75–96, `http://ideas.repec.org/a/eee/inecon/v54y2001i1p75-96.html`.

Hummels, David, Dana Rapoport and Kei-Mu Yi

1998 "Vertical specialization and the changing nature of world trade", *Economic Policy Review*, Jun, pp. 79–99, `http://ideas.repec.org/a/fip/fednep/y1998ijunp79-99nv.4no.2.html`.

Johnson, Robert C. and Guillermo Noguera

2012a "Accounting for intermediates: Production sharing and trade in value added", *Journal of International Economics*, 86, 2, pp. 224–236, http://ideas.repec.org/a/eee/inecon/v86y2012i2p224-236.html.

2012b *Fragmentation and Trade in Value Added over Four Decades*, NBER Working Papers 18186, National Bureau of Economic Research, Inc, http://ideas.repec.org/p/nbr/nberwo/18186.html.

Koopman, Robert, Zhi Wang and Shang-Jin Wei

2012 "Estimating domestic content in exports when processing trade is pervasive", *Journal of Development Economics*, 99, 1, pp. 178–189, http://ideas.repec.org/a/eee/deveco/v99y2012i1p178-189.html.

2014 "Tracing Value-Added and Double Counting in Gross Exports", *American Economic Review*, 104, 2 (Feb. 2014), pp. 459–94, http://ideas.repec.org/a/aea/aecrev/v104y2014i2p459-94.html.

Kummritz, Victor

2016 *Global Value Chains, Productivity, and Industrial Development*, CTEI Working Papers 2016-01, Centre for Trade and Economic Integration, Geneva.

Leontief, Wassily

1936 "Quantitative Input and Output Relations in the Economic System of the United States", *Review of Economics and Statistics*, 18, 3, pp. 105–125.

Quast, Bastiaan A. and Victor Kummritz

2015 *decompr: Global Value Chain decomposition in R*, CTEI Working Papers 2015-01, Centre for Trade and Economic Integration.

UNCTAD

2013 *World Investment Report 2013: Global Value Chains: Investment and Trade for Development*, tech. rep., United Nations Publication, Geneva.

Wang, Zhi, Shang-Jin Wei and Kunfu Zhu

2013 *Quantifying International Production Sharing at the Bilateral and Sector Levels*, NBER Working Papers 19677, National Bureau of Economic Research, Inc, http://ideas.repec.org/p/nbr/nberwo/19677.html.

4.A Output tables

country	*Average (i2e values)*	*Average (e2r values)*	*Average (i2e)*	*Average (e2r)*	*Δ 95-11 (i2e)*	*Δ 95-11 (e2r)*
arg	52,790	66,036	12.51%	15.65%	145.93%	30.04%
aus	178,117	343,084	13.35%	25.71%	18.21%	59.23%
aut	250,022	214,630	25.87%	22.21%	29.59%	39.90%
bel	437,578	285,355	32.53%	21.22%	10.66%	30.78%
bgr	51,393	19,864	38.01%	14.69%	32.70%	12.35%
bra	129,301	245,839	10.95%	20.83%	37.97%	57.77%
brn	2,412	20,438	4.51%	38.23%	-41.34%	103.27%
can	647,662	407,957	23.54%	14.83%	-3.54%	70.21%
che	334,258	343,657	21.84%	22.45%	23.45%	37.02%
chl	77,961	103,023	19.70%	26.03%	41.95%	42.06%
chn	1,831,434	1,293,766	24.07%	17.00%	62.57%	37.17%
col	21,746	57,971	9.12%	24.31%	-9.63%	93.42%
cri	21,400	11,671	28.07%	15.31%	25.42%	48.66%
cyp	12,327	8,448	22.01%	15.09%	0.27%	53.75%
cze	290,027	129,166	41.96%	18.69%	48.79%	11.22%
deu	1,640,838	1,628,409	22.51%	22.34%	71.81%	13.06%
dnk	224,697	165,653	29.42%	21.69%	38.07%	43.47%
esp	546,406	383,881	25.13%	17.66%	39.96%	35.86%
est	21,777	11,992	34.90%	19.22%	-3.58%	44.23%
fin	182,478	125,397	31.43%	21.60%	44.07%	6.99%
fra	888,006	773,925	23.01%	20.05%	44.76%	19.68%
gbr	766,576	909,659	19.52%	23.17%	25.71%	27.55%
grc	81,945	60,159	22.43%	16.47%	52.51%	51.65%
hkg	115,876	121,589	18.98%	19.92%	-7.57%	53.77%
hrv	20,725	13,277	20.09%	12.87%	-3.24%	-4.93%
hun	236,208	78,585	46.20%	15.37%	59.55%	23.83%
idn	116,161	238,302	12.97%	26.61%	-4.24%	96.09%
ind	356,692	298,471	21.34%	17.86%	178.40%	42.28%
irl	472,729	162,263	41.96%	14.40%	11.87%	19.69%
isl	11,301	8,977	29.32%	23.29%	84.40%	71.61%

country	*Average (i2e values)*	*Average (e2r values)*	*Average (i2e)*	*Average (e2r)*	Δ *95-11 (i2e)*	Δ *95-11 (e2r)*
isr	105,427	75,275	23.86%	17.04%	11.05%	53.19%
ita	778,367	641,040	23.21%	19.12%	53.48%	35.00%
jpn	582,907	1,388,524	11.95%	28.47%	164.46%	32.58%
khm	11,889	3,224	37.65%	10.21%	186.29%	-36.76%
kor	1,034,054	521,202	37.70%	19.00%	88.43%	18.61%
ltu	16,707	15,497	22.83%	21.17%	-4.03%	42.99%
lux	237,935	51,509	53.11%	11.50%	40.29%	-15.58%
lva	14,530	13,063	25.95%	23.33%	25.01%	34.38%
mex	479,806	214,457	28.82%	12.88%	20.84%	31.28%
mlt	14,762	4,931	43.26%	14.45%	-27.41%	108.31%
mys	517,084	215,738	41.45%	17.29%	33.36%	23.93%
nld	306,010	390,300	19.50%	24.87%	-14.29%	51.12%
nor	163,813	351,592	16.78%	36.01%	-13.46%	57.48%
nzl	39,712	33,912	17.08%	14.59%	-0.73%	48.61%
phl	103,838	82,783	29.04%	23.15%	-20.59%	101.26%
pol	273,027	198,877	29.34%	21.37%	99.99%	15.65%
prt	125,283	64,391	30.77%	15.81%	18.69%	37.51%
rou	59,385	54,956	24.23%	22.42%	14.95%	45.54%
rus	317,701	837,747	13.51%	35.62%	3.61%	56.47%
sau	57,392	547,987	3.95%	37.73%	-15.39%	56.02%
sgp	548,286	219,149	46.08%	18.42%	12.95%	61.94%
svk	140,548	61,023	44.95%	19.52%	47.60%	7.80%
svn	51,465	29,065	34.78%	19.64%	11.26%	58.26%
swe	355,353	262,980	29.09%	21.52%	8.68%	29.44%
tha	391,773	156,527	36.05%	14.40%	61.19%	23.84%
tun	35,124	18,724	30.17%	16.08%	30.65%	48.33%
tur	180,927	113,136	22.30%	13.95%	195.25%	12.71%
twn	649,797	353,241	39.52%	21.48%	41.83%	60.53%
usa	1,318,846	2,248,028	13.52%	23.04%	30.75%	26.87%
vnm	119,821	57,005	33.76%	16.06%	72.69%	19.95%
zaf	102,394	122,842	19.31%	23.16%	47.60%	24.98%

Table 4.9: GVC indicators by country

sector	*Average (i2e values)*	*Average (e2r values)*	*Average (i2e)*	*Average (e2r)*	Δ *95-11 (i2e)*	Δ *95-11 (e2r)*
c01t05agr	227,969	364,681	13.00%	20.79%	36.13%	27.51%
c10t14min	435,816	3,324,446	5.82%	44.38%	-4.51%	55.66%
c15t16fod	675,902	146,824	19.58%	4.25%	23.04%	32.73%
c17t19tex	679,185	174,555	23.30%	5.99%	9.49%	2.72%
c20wod	105,771	80,461	20.56%	15.64%	30.59%	63.12%
c21t22pap	302,255	344,112	19.23%	21.89%	27.63%	7.10%
c23pet	1,285,522	356,703	39.53%	10.97%	65.81%	-8.04%
c24chm	1,762,631	925,255	28.18%	14.79%	52.19%	-2.18%
c25rbp	446,528	309,253	27.89%	19.32%	38.26%	3.06%
c26nmm	161,574	126,800	22.09%	17.34%	41.44%	17.72%
c27met	1,340,507	868,120	31.03%	20.10%	36.53%	-11.26%
c28fbm	481,732	441,396	27.53%	25.23%	38.11%	6.86%
c29meq	1,398,864	570,006	26.50%	10.80%	39.73%	30.96%
c303233ceq	3,162,705	1,062,750	38.63%	12.98%	45.28%	18.74%
c31elq	713,928	301,827	30.71%	12.98%	37.32%	-3.42%
c34mtr	1,827,519	340,198	34.53%	6.43%	33.39%	7.25%
c35trq	745,063	203,834	30.54%	8.35%	40.89%	10.73%
c36t37otm	429,561	156,828	23.49%	8.58%	14.69%	66.57%
c50t52wrt	923,756	3,002,076	9.15%	29.75%	35.67%	33.76%
c60t63trn	1,293,729	1,507,876	17.88%	20.84%	60.51%	33.82%
c64ptl	81,529	322,823	12.39%	49.07%	85.71%	-7.06%
c65t67fin	371,026	992,980	13.51%	36.17%	85.73%	-8.65%
c71rmq	67,266	215,714	11.59%	37.18%	89.87%	13.54%
c72its	144,310	286,414	16.55%	32.86%	73.87%	-1.85%
c73t74bzs	386,997	1,922,031	11.27%	55.99%	44.41%	17.81%

Table 4.10: GVC indicators by sector

Final Remarks

Appendix A

`rnn`: Recurrent Neural Networks in `R`

Appendix B

Reproducibe Research Methods

Here I briefly discuss the combination of existing and new methods and tools that I used to try and make the research in this thesis as reproducible as possible.

First and moremost, it is essential to make clear which data is being used and in case it is primary data, how it was produced (e.g. research instruments). Where possible, the data itself should be included. If this is impossible to due e.g. privacy concerns of licencing issues, a clear procedure for obtaining the data should be documented. The data used in both South African studies is available upon request through an online portal. In the publicly available Git project (more details below), I include a description how to obtain the data used through the South African Labour & Development Research Unit .

Secondly, the computer code for producing the research results should be made available, in case this code produces intermediate data sets, where possible also make available.

Comment the code, use e.g. markdown type conventions, # for section, use ## for subsection. Add a header to the file containing (commented out):

1. file name
2. file purpose
3. author name
4. author email

In addition to describing the file purpose, the file name itself should also be meaningful. Typically the code in a file performs a certain action, as a rule, it can therefore best be described using a transitive verb, for instance. `import.R` the

resulting output (here the imported data) can then be saved using an intransitive verb as a file name, e.g. `imported.RData`.

Development versioning using Git, also useful for registering research designs (since logged), retracing steps. In addition to this, changes in all of the project files are logged using a version control software package. I use the open-source version control software called Git. Projects using version control such as this are called repositories. These repositories are published on the internet on websites such as GitHub or BitBucket.

Good reasons to use open-source R. Using open-source software in order to make the computational procedures of the statistical method verifiable. In my case this means that all statistical analysis has been done using the statistical programming environment R.

Additionally, when implementing new algorithms, packaging these as R libraries (extensions) and releasing under an open-source licence such as General Public Licence version 3 or later on the Comprensive R Archive Network (CRAN) website is a step towards replicablity. For instance, I packages the procedures implementing the Wang-Wei-Zhu algorithm used for the analysis as the `decompr` package. I uploaded this package to CRAN[1] and it has since been downloaded over 6,000 times[2]. This makes code useful for replicable research, since the original procedures, which may have contained hardcoded procedures (e.g. doing a loop 34 times, one for each country) now have to generalised and transformed into a function, since R packages can only contain functions. As a result of this, the code which we developed for the analysis of the TiVa data set, is now also being applied by other used to the wiod data set.

Adding weaving code . This also ties into `packrat`, since it makes clear which version of which libraries are used for the analysis.

[1] Available at: https://cran.r-project.org/web/packages/decompr/index.html

[2] Data from the RStudio servers only, other CRAN mirrors to not release statistics so the actual number is presumably higher. Note that this does not correspond directly to users, since updates require a new download.

Table B.1: R `sessionInfo()`

```
sessionInfo()

## R version 3.3.0 (2016-05-03)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.11.5 (El Capitan)
##
## locale:
## [1] C
##
## attached base packages:
## [1] methods   stats     graphics  grDevices utils     datasets  base
##
## other attached packages:
##  [1] decompr_4.1.0    lmtest_0.9-34    zoo_1.7-13       pglm_0.1-2
##  [5] maxLik_1.3-4     miscTools_0.6-16 printr_0.0.5     broom_0.4.0
##  [9] plm_1.5-12       Formula_1.2-1    magrittr_1.5     ggplot2_2.1.0
## [13] dplyr_0.4.3      knitr_1.13
##
## loaded via a namespace (and not attached):
##  [1] statmod_1.4.24     reshape2_1.4.1     splines_3.3.0
##  [4] lattice_0.20-33    colorspace_1.2-6   mgcv_1.8-12
##  [7] nloptr_1.0.4       DBI_0.4-1          fortunes_1.5-3
## [10] plyr_1.8.3         stringr_1.0.0      MatrixModels_0.4-1
## [13] munsell_0.4.3      gtable_0.2.0       bdsmatrix_1.3-2
## [16] codetools_0.2-14   psych_1.6.4        evaluate_0.9
## [19] labeling_0.3       SparseM_1.7        quantreg_5.24
## [22] pbkrtest_0.4-6     parallel_3.3.0     highr_0.6
## [25] Rcpp_0.12.5        scales_0.4.0       formatR_1.4
## [28] lme4_1.1-12        mnormt_1.5-4       digest_0.6.9
## [31] stringi_1.0-1      grid_3.3.0         tools_3.3.0
## [34] sandwich_2.3-4     cowsay_0.4.0       car_2.1-2
## [37] tidyr_0.4.1        MASS_7.3-45        Matrix_1.2-6
## [40] assertthat_0.1     minqa_1.2.4        R6_2.1.2
## [43] nnet_7.3-12        nlme_3.1-128
```

Piping functions to make to more intuitive. As will most lines will have a structure beginning with the object to which the output is assigned followed by the assignment operator, followed by the function, followed by the input data, followed by the argument. After which the following line again will begin with the output object, etc. We can see this for instance, if we use the built-in data set ‘swiss‘, which examines fertility levels in the French speaking regions of Switzerland in 1888. In this example I reproduce a simplified version of an original study this data, looking at the difference in fertiliity levels between predominantly Catholic and predominantly protestant agricultural regions.

Table B.2: swiss

```
# load data
data(swiss)

# inspect data
str(swiss)

## 'data.frame': 47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

The typical workflow is demonstrated below. We first define an *output* object for the filtered Agricultural regions, then the assignment operator, followed by the function, followed by the *input* data, followed by the argument. The second line then starts with a *new* intermediate object, etc. The final line starts with the function, then in last intermediate object, then the argument.

Table B.3: R without pipe

```
swiss_agr     <- filter(swiss, Agriculture > 50)
swiss_agr_grp <- group_by(swiss_agr, Catholic > 50)
summarise(swiss_agr_grp, mean(Fertility, na.rm=TRUE) )
```

Catholic > 50	mean(Fertility, na.rm = TRUE)
FALSE	65.62143
TRUE	79.51667

In this example I now use the `magrittr` package, which has the following pipe operator `%>%`. The first thing is the input data, followed by the transforming function, followed by the argument (criterium), this is passed to the next line where the second transforming function is the first item, follow by its respective argument, this is passed on to last line, where the data is summarised, in this case by computing the mean.

Table B.4: R with pipe

```
swiss %>%
        filter(Agriculture > 50) %>%
        group_by(Catholic > 50) %>%
        summarise( mean(Fertility, na.rm=TRUE) )
```

Catholic > 50	mean(Fertility, na.rm = TRUE)
FALSE	65.62143
TRUE	79.51667