Spatial heterogeneity, generational change and childhood socioeconomic status: microeconometric solutions to South African labour market questions

by

Dieter von Fintel

Dissertation presented for the degree of

Doctor of Philosophy (Economics)

in the Faculty of Economic and Management Sciences

at the University of Stellenbosch

Supervisor: Dr Rulof Burger

Co-supervisor: Prof Servaas van der Berg

December 2014

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

With regard to Chapter 3, the nature and scope of my contribution were as follows:

Nature of contribution	Extent of contribution (%)
Cleaning and analyzing data; estimation; write-up of literature review and analysis of results.	75%

The following co-authors have contributed to Chapter 3:

Name	E-mail address	Nature of contribution	Extent of contribution (%)
Dorrit Posel	_____	Helped formulate the research question and focus, revised introduction and literature review, reviewed and commented on drafts of the chapter.	25%

Signature of candidate: ……

Date: 24 November 2014……..

Declaration by co-authors:

The undersigned hereby confirm that

1. the declaration above accurately reflects the nature and extent of the contributions of the candidate and the co-authors to Chapter 3,
2. no other authors contributed to Chapter 3 besides those specified above, and
3. potential conflicts of interest have been revealed to all interested parties and that the necessary arrangements have been made to use the material in Chapter 3 of this dissertation.

Signature of co-author:

Date: 5 August 2014

Abstract

Microeconometric techniques have improved understanding of South Africa's labour market substantially in the last two decades. This dissertation adds to this evidence by considering three separate labour market questions, with particular attention to data quality and the application of credible methodology.

Firstly, wage flexibility is investigated. Whereas selected previous microeconometric evidence suggests that wage setters in South Africa are highly responsive to external local labour market circumstances, it is not corroborated by macroeconomic and other microeconometric studies. This question is interrogated again, with particular attention to methodological issues in wage curve estimation. The latter is a robust negative relationship between individual wages and local unemployment rates, found in most countries, except where bargaining is highly centralized. Adding time variation to the data allows controls for spatial heterogeneity to be introduced, leading to the conclusion that wages are really inflexible in the short-run. Rather, the trade-off between wages and local unemployment that previous work has found represents a long-run spatial equilibrium. This finding is robust to instrumentation for reverse causality and the measurement error that is associated with choosing incorrect labour market demarcations.

Secondly, the reliability of retrospective data related to childhood is investigated, with the view of estimating the long-run influence that early life circumstances have on adult outcomes. Two indicators, parental education and subjective rankings of childhood socioeconomic status, are evaluated. The first set of indicators has poor response rates, as many South African children live without their parents. Where respondents do volunteer this information, they answer consistently across waves. Subjective rankings have higher response rates, as they require respondents to provide information about their own past, and not about those of their parents. However, individuals' assessments are inconsistent over time, despite being asked about the same point in the life cycle. They tend to change their view of the past in line with adjustments to perceptions of their position in the village income distribution and subjective well-being, providing clear evidence of anchoring. Instrumental variables analysis has been used in previous studies to account for measurement error in subjective data. However, if anchoring affects all assessments of the past and potential outcome variables (such as employment), microeconometric techniques will yield biased estimates of the effects of childhood on long-run outcomes.

Finally, age-period-cohort models for South African labour force participation are estimated. This chapter is the first contribution to relax the assumption that cohort differences must remain permanent over the life cycle. Monte-Carlo simulation studies show that highly interactive specifications can partially recover the true underlying process. Using a variety of techniques (imposing behavioural restrictions and atheoretical approaches), this study shows that cohort effects in labour force participation can be temporary in South Africa, though more data is required to verify this conclusively. Regardless of technique, a distinct surge in labour force participation is noted for the group born after 1975. Pertinently, the combination of testable assumptions and highly flexible estimation can yield credible age-period-cohort profiles, despite the many disputes noted in the literature. Previous evidence of a surge in participation for the post-1975 cohort can now be shown to be temporary rather than a part of a long-run generational increase.

Opsomming

Mikro-ekonometriese tegnieke het kennis oor die Suid-Afrikaanse arbeidsmark aansienlik uitgebrei in die afgelope twee dekades. Hierdie proefskrif dra by tot hierdie bewyse deur drie afsonderlike arbeidsmark vraagstukke te beskou, met die klem op datagehalte en toepassing van geloofwaardige metodologie.

Eerstens word die kwessie van loonaanpasbaarheid beskou. Waar sekere vorige mikro-ekonometriese bewyse aandui dat loonbepalers in Suid-Afrika sterk op eksterne plaaslike arbeidsmarktoestande reageer, word hierdie bevinding nie deur makro-ekonomiese en ander mikro-ekonometriese studies ondersteun nie. Hierdie vraag word dus opnuut ondersoek, met die klem op metodologiese kwessies wat ‘n invloed op die beraming van die loonkurwe het. Laasgenoemde is die negatiewe verhouding tussen individuele lone en plaaslike werkloosheidskoerse wat in die meeste lande geld, behalwe daar waar loonbedinging sterk gesentraliseer is. Deur tydsvariasie by die data te voeg, is dit moontlik om vir heterogeniteit oor ruimte voorsiening te maak, wat tot die gevolgtrekking lei dat lone inderdaad onbuigsaam oor die korttermyn is. Die afruiling tussen lone en plaaslike werkloosheidskoerse wat vorige navorsing bevind het, verteenwoordig eerder ‘n langtermyn ruimtelike ewewig. Hierdie bevinding is nie sensitief vir instrumentasie nie. Laasgenoemde is nodig om voorsiening te maak vir moontlike sydigheid wat kan ontstaan indien die rigting van kousaliteit omgekeerd is, sowel as metingsfoute wat daarmee gepaard gaan as navorsers die plaaslike arbeidsmark verkeerd definiëer.

Tweedens word die betroubaarheid van data wat volwassenes vra om hulle kinderomstandighede te onthou, ondersoek. Die uiteindelike doel is om vas te stel of omstandighede vroeg in die lewe ‘n invloed op die uitkomstes van volwassenes het. Twee veranderlikes, naamlik ouers se opvoedingsvlakke en die subjektiewe terugskouende sosioekonomiese rang in respondente se kinderdae, word geëvalueer. Die eerste stel veranderlikes is onderhewig aan lae reaksiekoerse omdat ‘n aansienlike hoeveelheid Suid-Afrikaanse kinders sonder een of beide ouers grootword. Waar respondente wel hierdie inligting verskaf is individue se antwoorde konsekwent tussen twee golwe van ‘n paneelopname. Die vraag na die subjektiewe rang lewer beter reaksiekoerse omdat dit vereis dat respondente inligting oor hulle eie verlede verskaf, en nie oor dié van hul ouers nie. Nietemin is individue se antwoorde strydig oor tyd, ten spyte daarvan dat hulle inligting oor dieselfde tydstip in die lewenssiklus moet verskaf. Hulle is geneig om hulle opinies oor die

verlede in lyn met veranderende persepsies van hul huidige posisie in die dorpsinkomsteverdeling, sowel as hulle eie subjektiewe welstand, aan te pas. Dit verskaf dus ‘n sterk aanduiding dat mense hulle antwoorde oor die verlede in huidige toestande anker. Instrumentele veranderlike analise is in vorige studies aangewend om voorsiening te maak vir metingsfoute in subjektiewe data. Indien inligting oor die verlede, asook moontlik uitkomsteveranderlikes (soos indiensname), geanker word in huidige persepsies, sal mikroekonometriese tegnieke egter steeds sydige beramings van die impak van kinderdae op langtermyn uitkomstes bied.

Laastens, word sogenaamde ouderdom-periode-kohort modelle op Suid-Afrikaanse arbeidsmarkdeelname data toegepas. Hierdie hoofstuk is die eerste bydrae wat die aanname dat kohortverskille permanent moet bly oor die lewenssiklus laat vaar. Monte-Carlo simulasies dui aan dat hoogs interaktiewe spesifikasies die onderliggende proses gedeeltelik kan weerspieël. Verskeie tegnieke word aangewend (insluitend dié wat gedragsaannames afdwing asook ateoretiese benaderings) wat wys dat kohorteffekte in arbeidsmarkdeelname tydelik kan wees. Tog word meer data benodig om hierdie stelling sonder twyfel te bevestig. Onafhanklik van die tegniek wat gebruik word, is dit duidelik dat ‘n skerp toename in arbeidsmarkdeelname plaasgevind het vir die groep wat na 1975 gebore is. Verder is dit beduidend dat die kombinasie van toetsbare aannames en hoogs buigsame beramers ‘n geloofwaardige oplossing vir die ouderdoms-periode-kohort probleem verskaf, ten spyte van die vele twispunte wat in die literatuur uitgelig word. Vorige bewyse van ‘n toename in arbeidsmagdeelname vir die post-1975 kohort kan nou as ‘n tydelike tendens bestempel word, eerder as ‘n deel van die langtermyn toename oor generasies.

Acknowledgements

I am grateful to my supervisors, Dr Rulof Burger and Prof Servaas van der Berg, who provided valuable guidance in refining the ideas contained in this dissertation. They have been continuously supportive of my work and acted in ways to promote my development as a young researcher, always being open to engage and collaborate. I would also like to thank Dr Francis Teal (who supervised me as a visiting student at Oxford University) and Prof Dorrit Posel (who mentored me as part of Stellenbosch's Mellon Early Researcher Career programme), for their substantial inputs and for challenging me to think differently about my research. Prof Andrie Schoombee, chair of the Department of Economics at Stellenbosch University, has been generous in clearing extended periods of my teaching schedule to focus on my dissertation. The environment created by the department is conducive for young researchers to grow, and I am thankful that I may work in such a collegial and stimulating setting.

I thank my friends and colleagues in the Research on Socioeconomic Policy and African Economic History groups at Stellenbosch, as well as fellow students at the Centre for the Study of African Economies at Oxford, for lively discussions, which have sharpened me as a researcher. The Commonwealth Scholarship Commission made it financially possible for me to spend one academic year at Oxford University, and the Mellon Early Researcher Career programme at Stellenbosch University funded academic mentorship during the time of writing my dissertation. REDI3x3 supported parts of this research financially. I also acknowledge the support of the Institute for the Study of Labour (IZA), which allowed me to attend their Young Scholars' Programme in Washington, DC. This useful opportunity enabled contact with prominent researchers. Earlier versions of this research were presented at the annual Centre for the Studies of African Economies Conference in Oxford, the weekly research workshop of the same institution, the Gorman student seminar at Oxford, the African Economic Conference, the Microeconometric Analysis of South African Data conference, the ReSEP brownbag lunches and workshops, the Stellenbosch Department of Economics seminar and various Economic Research Southern Africa workshops. I appreciate the many useful comments of participants. I thank Dr Surette Bierman for statistical advice, and Garth Stephenson and Ilze Boonzaaier for GIS assistance.

I am grateful to many friends who stood by me during the completion of this dissertation. Johan Fourie and Louw Pienaar have co-authored other papers with me that have inspired me to pursue better research and to see the bigger picture. Many have offered continuous words

of encouragement: Xander Kritzinger, Johann van Eeden, Leandro and Carmen Boonzaaier, Keith and Jacobeth Whiting, Mahieu and Alanna van der Linde, Bianca Bottega, Hein Gerber, Rousseau Lötter, Matthys and Elsje Saayman (who also assisted in transcribing data). I also thank personal mentors who generously shared their wisdom with me: Petrus and Belinda Beukes, George Malek and Donald Hay.

Marisa Coetzee has spent many hours listening to me talk about my research and just about anything else, always challenging me to think more clearly. She has been immensely patient and has been my biggest supporter. I cherish the love and sincerity that she and her family show me.

My own family has been unceasing in their love and support, and often believed in me more than I did myself. I thank my parents, Peter and Helga von Fintel for sacrificing so much for me to become the person who I am, and for being dedicated to all their children, regardless of the circumstances. My brothers and sisters, Bernd and Helene Peters, Rudi and Caryn von Fintel, have been gracious to me and loved me despite my weaknesses. My late grandfather, Prof Wilhelm Maré, and our cousins, Werner and Empie Krull, have been role models of excellence for me.

To God alone be the glory.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and research questions

Since the first nationally representative household survey was enumerated in South Africa in 1993, microeconometric research on the labour market has expanded substantially[1]. While data availability is a vital enabling factor, interest in the South African labour market has been generated primarily by unique features that distinguish it from many others. The legacies of separate spatial and racial development policies, skills mismatches and apparently high labour market rigidity are factors that contribute to South Africa being one of the highest unemployment economies globally. The very first objective listed in the National Development Plan is to reduce unemployment, while the other aforementioned contributing factors are also given priority (Republic of South Africa, 2012). However, post-apartheid social change has been slow, with low levels of mobility within and across generations remaining (Piraino, 2014; Finn et al., 2014; Lechtenfeld & Zoch, 2014), and the labour market playing a dominant role in high levels of overall income inequality (Leibbrandt et al., 2009). Microeconometric methods are therefore vital to monitor the progress of social upliftment, and have already contributed substantially to the policy discussion.

This dissertation focuses on selected labour market issues, with the emphasis on data quality and the implementation of credible microeconometric methodology to arrive at useful policy conclusions. In most instances alternative solutions are presented, in order to verify or challenge existing knowledge, and to address existing concerns in estimation. However, the dissertation also warns of potential problem areas that remain in using data effectively. Chapter 2 interrogates the issue of wage flexibility, illustrating why studies that do not account for spatially persistent differences yield misleading conclusions. This is particularly relevant because of the strong spatial segmentation that characterises South Africa's labour market. In addition, the study shows that local labour markets are not necessarily as small as many authors assume them to be. Previous evidence is rebutted, concluding that wages are generally not set in relation to local labour market conditions. Hence, the labour market is inflexible along this dimension. Chapter 3 turns directly to the issue of long-run mobility, though the focus falls on evaluating the reliability of data. Life course data, which is used to establish the socioeconomic conditions of households during childhood, are rarely available to researchers. These indicators are required to assess whether adults' labour market prospects

[1] Fourie (2011) reviews a large section of this literature.

are rooted in childhood backgrounds. Should childhood circumstances have far reaching implications for adults, it places the emphasis on early intervention rather than only finding solutions to (inter alia) labour market problems once individuals are of working age. However, this study illustrates that while individuals' recall of their past can potentially fill the data gap, current circumstances strongly influence South Africans' assessment of the past. Researchers should be aware of the contamination of these measures, and how they affect estimates of life course mobility. Chapter 4 returns to the issue of generational increases in labour force participation, which have in turn contributed to the rise in unemployment. Knowing whether changes in the labour market are permanent or temporary is essential in order to understand the future. Labour market entry has occurred at progressively earlier ages among recent generations, which in turn has fuelled youth unemployment (Branson & Wittenberg, 2007; Burger & von Fintel, 2014). Should these higher rates of participation be generation-specific, they may persist across the life cycle and present a permanent unemployment problem if individuals are also not absorbed into jobs over this horizon. However, if the phenomenon is specific only to the early part of the life cycle, and for a specific generation, interventions should be targeted at only this limited group and may only need to be temporary (which is potentially the case for the recently implemented youth wage subsidy (Levinsohn et al., 2014)). Age-period-cohort decompositions have been applied in a broad literature from sociology to epidemiology, to disentangle generational from life cycle components in outcome variables. While this methodology is conceptually simple, identification has challenged researchers from all disciplines. Furthermore, the additive nature of the model prevents results from having heterogeneous life cycle effects, which is necessary to answer the policy question proposed here. Multiple identification strategies, along with extensions to include heterogeneous solutions, are investigated for their reliability in drawing conclusions about the distinction between permanent and temporary labour force participation patterns in South Africa.

The rest of this introductory chapter first sketches selected features of the South African labour market. They contextualise the questions that will be investigated in the later chapters of this dissertation. Thereafter, each of the scenarios is discussed with reference to the microeconometric solutions that this dissertation investigates.

1.1. Labour market background

South Africa's labour market has evolved substantially in the last few decades, leading to a situation that is unique among developing countries. Its unemployment rate is among the highest in the world, while many other African labour markets are characterised instead by underemployment (Fields, 2011). Since the fall of apartheid, unemployment continued to rise,

peaking in 2003 and never returning to lower levels than those that prevailed in the early 1990s. Banerjee et al. (2008) conclude that this change does not represent a temporary adjustment after democratisation, but is rather the product of long-run shifts that have led to a sustained high unemployment equilibrium. A sectoral shift emerged from the 1970s, whereby employment in primary industries declined relative to a great expansion in tertiary sector occupations (Bhorat & Hodge, 1999). This has contributed to high structural unemployment, as the skills composition of labour force participants has not altered to match these changes in labour demand. While this phenomenon of skills-biased technological change is not specific to South Africa, other long-run changes have added to this trajectory. During a similar period (even before the fall of apartheid) previously prohibited black unions were unbanned, allowing wages of the unskilled to rise substantially. Employment creation for unskilled workers slowed down concurrently (Lewis, 2001). The strong influence of labour unions has continued into the post-apartheid era, with high wage premia for workers that are covered by collective bargaining agreements (Magruder, 2012; Bhorat et al., 2012). Furthermore, these agreements are associated with substantial employment losses. The result is that wage growth has exceeded productivity growth – unemployment rose, though the increase in joblessness has not stopped high wage demands from continuing (Fedderke, 2012; Klein, 2012).

In addition to political changes, social and demographic shifts emerged, with a gradual feminization of the labour force and a surge in entry for individuals born after 1975 (Casale & Posel, 2002; Burger et al., 2014). The latter shift occurred due to changes in education policies that disallowed learners (who had not yet passed sufficient numbers of grades by specified ages) to continue with their schooling. Instead of continuing their schooling in Further Education and Training (FET) colleges, many entered the labour market. In the face of the labour demand constraints mentioned above, these large long-term increases in participation have contributed to rising unemployment. Banerjee et al. (2008) again emphasize that the growth in female labour force participation is unlikely to reverse due to new social norms, so that social change has indirectly contributed to unemployment. However, it is not certain whether the pattern among recent generations is also permanent in nature: could we expect participation rates to stabilise or reverse somewhat as individuals from affected cohorts age, alleviating pressure on the large stock of unemployment? Banerjee et al. (2008), however, suggest that South Africa's labour market is unlikely to escape from its high unemployment equilibrium through self-adjustment, but instead requires intervention.

1.2. Wage flexibility and spatial heterogeneity

Despite assertions that the labour market does not readily self-correct in response to high unemployment, Kingdon & Knight (2006a) do find that wage flexibility (in response to slack

labour market conditions) is as high in South Africa as in developed, low unemployment economies. These findings raise the question whether the labour market does in fact have the ability to self-adjust for large labour surpluses through wage moderation. The collective evidence presented above suggests that this has not been the case in reality. In addition, macroeconomists would not agree, maintaining that wages are downwardly rigid, with sustained growth even in economic downturns (Fedderke, 2012; Klein, 2012). Fourie (2011) emphasizes that labour market researchers from the macro and micro traditions do not interact sufficiently with each others' work. In this particular case, macroeconomic studies show the opposite to this one piece of microeconometric evidence, calling for a re-evaluation of wage flexibility. Other microeconometric studies (for instance, Magruder, 2012) also do not support the notion of wage flexibility. In light of Kingdon & Knight's (2006a) own surprise that South Africa (as unique as it is) apparently conforms to an international "empirical law of economics" (Blanchflower & Oswald, 2008), wage curve estimation will be re-evaluated in Chapter 2. Wage curves estimate the relationship between individual's wages and local unemployment rates for the regions in which they live. Most economies' wage curve elasticities are negative. When unemployment is high, wages can serve as a moderating factor for surplus labour by falling. If the relationship is absent or positive, the labour market can be considered inflexible, and even small labour surpluses are unlikely to be cleared through slower wage growth.

Chapter 2 of this dissertation addresses the discrepancy between existing microeconometric and macroeconomic evidence, suggesting that previous results confound long-run and short-run phenomenon. Research in other countries has shown that it is important to account for spatial heterogeneity in wage curve analysis, especially when bargaining is centralized (Albaek et al., 2000; Daouli et al., 2013). Magruder (2012) illustrates the importance of centralized collective bargaining for wage setting in South Africa. Additionally, his empirical labour market models emphasize the role of accounting for spatial heterogeneity in South African data. Because of the availability of a longer time period of labour market data, it is now possible to account for spatial heterogeneity more adequately, a solution that was not possible using the cross section data that Kingdon & Knight (2006a) did.

In addition, Chapter 2 addresses a concept that is not formally investigated by earlier research. Given that wage curve estimates are reliant on "local" unemployment rates, most researchers use geographic demarcations that are immediately available in survey data. Some researchers suggest that the smaller the demarcation, the closer it is to approximating "local" conditions. However, where wage setting factors are uniform across large areas (such as when minimum wages and collective bargaining agreements cover multiple districts), "functional" labour markets may be much larger than initially assumed. The chapter illustrates that

choosing the incorrect demarcation for wage curve analysis biases results towards zero, thereby favouring conclusions that would suggest inflexibility. This occurs whether labour markets are chosen to be either too small or too large compared to the optimum. In most existing studies internationally this may be inconsequential, as they find large negative elasticities. However, in instances where wages are only moderately flexible, it is possible that one could conclude that they are rigid if one chooses a sub-optimal local labour market definition. Instrumental variables estimates are investigated to provide a solution to this problem.

Finally, Chapter 2 also highlights that the standard procedure that accounts for reverse causality and measurement error in wage curve analyses is not satisfactory when unemployment is time persistent and spatial heterogeneity is simultaneously important. Time lagged instruments for unemployment are weak when spatial fixed effects remove the large persistent component of unemployment. This study proposes using spatially lagged instruments. While the latter instruments are strong, they are potentially not exogenous, but present a robustness check to estimates that only account for spatial heterogeneity. Best estimates, accounting for spatial heterogeneity, optimal labour market size and reverse causality, conclude that wages are inflexible in South Africa, and do not respond to high local unemployment.

1.3. Childhood reach and recall data

An emerging literature in South Africa considers income mobility within households, either over short time periods (Finn et al., 2014; Lechtenfeld & Zoch, 2014) or across generations (Piraino, 2014). However, there is growing interest internationally to study long-run mobility over individuals' lifetimes, and for outcomes beyond income (Almond & Currie, 2011; Cunha et al., 2006). In particular, can poor children live as well-off as rich children once they reach adulthood? Are gaps in welfare dependent on early life circumstances, and are interventions necessary in early childhood to prevent the same gaps from persisting later in life? Various studies consider effects on school attainment and performance, health or obtaining jobs in adulthood.

International studies on the reach of childhood most often rely on long-run cohort studies, which are rarely available in the South African context[2]. The alternative is to use individuals' recall assessment of childhood circumstances as explanatory variables for adult outcomes. Chapter 3 assesses whether these indicators are suitable for this type of research. In particular,

[2] One exception is the Birth to 20 data that follows an urban South African sample that was born in 1990. This data is not publicly available, hence few economists have worked on childhood reach in South Africa. An isolated example of such work is conducted by Casale et al. (2014).

responses to various retrospective questions are compared in two waves of panel data. Objective measures of childhood circumstances ask children to recall their parents' education levels. These measures are compared to questions that ask respondents to rank their socioeconomic status at age 15 on a ladder with six rungs. Should responses that relate to the same point in individuals' life cycles be largely consistent over time, then the data is likely to be reliable. However, as chapter 3 illustrates, each of these indicators is poorly recalled in South Africa. Firstly, adults have low response rates on retrospective questions, especially those that ask respondents to provide information about their parents. One reason is that many South Africans never lived with both biological parents during childhood, because of the migrant labour system and the low marriage rates among the black sub-population. Furthermore, it is possible that this indicator is more suitable for measuring intergenerational mobility rather than lifetime mobility: it directly represents parental endowments, while only indirectly pointing to childhood circumstances resulting from parental human capital. Additionally to selective response, current socioeconomic circumstances result in anchoring of the past onto the present (Tversky & Kahneman, 1974), especially for the subjective ranking of socioeconomic status. Hence, a trade-off arises: the subjective ranking elicits high response rates but with poorer consistency over time; objective recall of parental education has low reporting rates, but greater consistency across time among respondents.

Neither measure may therefore be entirely suitable to conduct estimates of long-run mobility. Take the case of anchoring in the subjective ranking. Current shocks to outcome variables (such as employment in adulthood) also influence respondents' recollection of the past. Consequently, an artificial relationship arises if these indicators are regressed against each other, based purely on errors in perception. While repeated measures of the same indicator (using panel data) could potentially serve as instruments for each other, measurement errors must be uncorrelated across time for this to be a valid identification strategy. However, persistent anchoring errors preclude this type of causal analysis. As a result, alternatives are required to credibly use retrospective data to model life course and intergenerational mobility.

1.4. Generational labour force participation and heterogeneity in age-period-cohort analyses

As noted before, labour force participation has experienced a long-run rising trajectory in South Africa. In particular, females have progressively entered the labour market, as their education levels have risen, marriage rates have declined and they have become more likely to head households (Casale & Posel, 2002). While it is easy to conceive of such a movement as a permanent generational rise in participation that will not reverse, increases for other demographic groups may not share this property. In particular, previous research that has

focussed on additively separating labour force participation into life cycle, business cycle and generational components, shows that black individuals born after 1975 entered the labour market in substantially greater proportions than earlier generations (Burger & von Fintel, 2014). Will this distinct cohort behaviour disappear, with labour force participation resembling that of other generations once they age? In other words, it may be possible that only a temporary cohort effect exerted influence early in the relevant generation's life cycle. If this is the case, this particular cohort, that is also prone to higher levels of unemployment, may also only need temporary policy assistance. In contrast, if the cohort effect is permanent, it is likely to have far reaching consequences, with targeted interventions required throughout the life cycle. Hence, what some might observe to be a "youth" unemployment problem may rather be a problem of a specific generation. Such a situation would require substantially different policies to solve, compared to when the youth of all generations must be targeted.

The estimators that separate these effects assume *a priori*, however, that cohort components are permanent, and that differences across generations are constant at all points along the life cycle. Hence, substantial heterogeneity is ignored if these assumptions are not met in reality (Glenn, 1976). Age-period-cohort estimators are highly debated in the literature, even without relaxing the assumption of permanent cohort effects: various methods attempt to solve for the fact that the three components are perfectly collinear, but nevertheless represent distinct phenomena.

Chapter 4 assesses, using a Monte-Carlo simulation study, whether any of the array of estimators adequately recovers the age-period-cohort components from a simulated data generating process that resembles South Africa's labour force participation trajectory. This simulation study is the first to also relax the assumption of a permanent cohort effect, by allowing for an interactive age profile: the result is that some cohort affects can only be present in parts of the life cycle. The generated process resembles a youth surge in participation for one group of labour market entrants that subsequently slows down and normalizes later in life. In addition to some of the standard estimators proposed in the literature, adaptations of these to include interactive components are explored. The latter are highly flexible, borrowing from non-parametric innovations in the statistical literature (Hastie & Tibshirani, 1990), and also adapting the prominently used Intrinsic Estimator (Yang et al., 2004) to include fully interactive solutions. In addition, testable restrictions based on the fact that second derivatives of profiles are unique, are also extended to the interactive case (McKenzie, 2006). After establishing the effectiveness of each estimator, the interactive solutions are applied to South African labour force participation data. The analysis confirms that a portion of cohort effects (over and above a long-run trend) may be non-permanent, and that some of the generational rise in participation is likely to slow down later in the life cycle.

However, additional data from later periods should be monitored to confirm or refute these changes, prompting future research in this area.

1.5. Summary

The research questions contained in this dissertation focus on policy questions relating to South Africa's labour market. Microeconometric solutions and outstanding problems are presented in each instance. Firstly, while existing microeconometric evidence on labour market flexibility contrasts with the literature from the macroeconomic strand, Chapter 2 illustrates that this discrepancy arises because of the inability of cross section data to adequately identify these effects. Wage curves that fully account for spatial heterogeneity, appropriate local labour market boundaries and reverse causality reveal that South African wage setters are insensitive to labour market conditions. Secondly, Chapter 3 assesses whether retrospective data can fill gaps in the knowledge of life course mobility and childhood reach. Anchoring in recall assessment of childhood socioeconomic status potentially prevents this possibility without further research. Finally, Chapter 4 adapts age-period-cohort estimators to account for potentially non-permanent generational effects. Simulation studies and applications to South African labour force participation data highlight the importance of acknowledging this type of heterogeneity. Assuming additive components in age-period-cohort components poses the danger that policies that should only apply during a section of a generation's life cycle are incorrectly extended throughout these groups' lives.

The microeconometric techniques applied in each chapter highlight that standard approaches must be reconsidered to arrive at correct conclusions about labour market phenomenon. This dissertation contributes to the exiting literature by making recommendations on how to credibly build various models to understand South African labour market conditions. It illustrates shortcomings in past estimates, and provides means to make better policy conclusions.

Chapter 2

Wage flexibility in a high unemployment regime: spatial heterogeneity and the size of local labour markets

2.1. Introduction

South Africa's economy consistently ranks among the lowest internationally when it comes to firms' discretion about wages that they pay their employees (World Economic Forum, 2014). Rigid labour laws and strong unions are regularly cited as constraints to employment creation, with wages growing faster than labour productivity (Fedderke, 2012; Klein, 2012; Banerjee et al., 2008). Despite high structural unemployment, wage demands have not moderated, even during downturns in economic activity. Collective bargaining agreements have also contributed to this situation, with substantial wage premia that limit job creation capacity, especially among smaller firms (Magruder, 2012). Legally, wages that are negotiated between groups of firms and unions can be extended to entire industries, even to firms with only non-unionized workers. As a result, many firms have little discretion with regards the wages that they pay, with wage setting far removed from an individual worker-firm match.

However, annual rankings of wage flexibility are based on subjective assessments, with little concrete evidence to support them. One piece of microeconometric research that measures flexibility in wage setting surprisingly concludes that South African workers behave similarly to those in the rest of the world: when local labour market conditions are slack, wage demands are apparently moderated, so that firms are able to offer lower wages in less favourable economic climates. Kingdon & Knight (2006a) estimate wage curves using cross section data, and conclude that the elasticity of individual wages with respect to local unemployment rates is close to -0.1, a widely cited figure that is established for most developed economies. This result is so consistent across studies that it has been termed an "empirical law of economics" by its originators (Blanchflower & Oswald, 2008). In high unemployment economies it is unlikely, however, that wages can drop sufficiently to clear labour surpluses. Wages may already be close to a lower bound to enable subsistence, so that downward adjustment is not possible, even if unemployment grows. Additionally, in other countries with high degrees of centralized bargaining, wages have been found to be insensitive to high local unemployment (Albaek et al., 2000; Daouli et al., 2013).

The results of Kingdon & Knight (2006a) therefore contrast with two commonly understood regularities about the South African labour market: that wages are sticky downward in response to the protection of workers' interests, and that when unemployment is as high and structural as it is, that wages are not a primary mechanism by which the labour market imbalance is corrected. The authors also express their surprise that despite multiple robustness checks, their results conform to those of international labour markets, rather than one that faces the rigidities that South Africa does.

This chapter questions whether wage flexibility is as high in South Africa as existing estimates have suggested, or whether it instead conforms to reports of rigidity in macroeconomic research (Fedderke, 2012; Klein, 2012) and other microeconometric studies (Magruder, 2012). It builds on previous literature to provide a concrete metric – wage curve elasticities – by which to evaluate whether wage setting does respond to labour market conditions, especially where they are severely slack. A first objective is to highlight the importance of spatial heterogeneity in the estimation of wage curve elasticities. International evidence suggests that this is particularly important when centralized bargaining is pervasive (Albaek et al., 2000). Regional unemployment may have persistent components (as determined by institutions such as separate development and collective bargaining), which should be modelled with spatial fixed effects. This framework allows for the separation of short-run from long-run wage adjustments to local labour market conditions, a possibility not yet considered in the South African context. Previous estimates were unable to account for spatial heterogeneity effectively, due to the lack of time variation in survey data. Additional periods of Labour Force Survey (LFS) data allow for these estimates to be updated, fully accounting for spatial heterogeneity.

Secondly, this paper highlights the consequences of estimating wage curves with demarcations that do not naturally constitute functional local labour markets. Most authors assume that smaller regions better represent local labour market boundaries. However, it is (for instance), possible that when institutions such as bargaining councils cover multiple small regions, that larger geographic areas naturally integrate. Given the availability of various demarcations in the LFS data, this paper establishes the best approximation for "natural" labour markets in South African surveys. Importantly, wage flexibility can be vastly underestimated if regions that are either too small *or* too large are chosen to conduct estimation. Finally, wage curve estimates usually solve for reverse causality and measurement error by instrumenting with time lagged unemployment rates. Data limitations again precluded this approach in previous studies for South Africa. However, this study shows that this instrument is weak when spatial heterogeneity plays a central role in estimation, and fixed effects remove common variation across time. As an alternative, *spatial* lags are

proposed, exploiting geographic variation in unemployment as an instrument. Accounting for each of spatial heterogeneity, reverse causality and potentially incorrect labour market demarcations, estimates confirm that South African wage setters are inflexible in their response to slack labour market conditions. These results suggest that the data limitations faced by Kingdon & Knight (2006a) led to conclusions that were at odds with what macroeconomic and other microeconometric research has shown to be a reality. This contribution emphasizes the importance of addressing various methodological concerns in a high unemployment region with high degrees of spatial heterogeneity.

The rest of this chapter is structured as follows. Section 2.2 briefly interrogates the literature that links local labour markets and spatial heterogeneity. Section 2.3 surveys the wage curve literature, with a specific focus on the conditions under which wage inflexibility is measured: both empirical elements (such as the optimal radius by which to measure labour markets and spatial heterogeneity) and institutional rigidities are considered. Section 2.4 introduces the dataset and empirical strategy that have been adopted, while section 2.5 outlines a full set of results. Section 2.6 concludes.

2.2. Spatial heterogeneity and local labour markets

Wage flexibility estimates require the identification of spatial units that constitute local labour markets. These geographic units are determined by various equilibrating economic forces, notably the ability of labour and firms to move across regions, in order to find the best opportunities. Hence, the review first focuses on how various spatial equilibria arise and/or persist, with specific references to South Africa. Thereafter, the reach of local labour markets is explored in light of labour market institutions that are regionally defined. Each of these sub-sections informs the discussion that follows.

2.2.1. Spatial equilibrium

South Africa has a segmented labour market along multiple dimensions: barriers to entry into the informal sector raise unemployment (Kingdon & Knight, 2004), the apartheid migrant labour system separated individuals over space (Posel, 2010), as did segregation within cities and job reservation within firms (Mariotti, 2012). While some of these characterisations are to a degree specific to the current situation, labour markets are rarely homogenous across space within most countries (Moretti, 2011): in equilibrium, regions may have vastly different productivity levels, wages and capacity for employment creation. This contrasts with the product market, where the law of one price predicts that arbitrage will lead to equilibrium with fairly uniform product prices across space. Potentially equivalent equilibrating forces in the labour market come in the form of migration of workers towards regions offering high

wages and where favourable conditions entail a high enough probability of finding a job to warrant a move. Similarly, firms may find it profitable to locate where wages are low, productivity is high and hence create employment.

Nevertheless, wages and employment levels rarely reach a point of uniformity across regions, despite movements of people that should achieve this. For instance, the seminal contribution of Harris & Todaro (1970) postulates why urbanization in developing countries typically occurs despite high unemployment levels there. Their two-sector model suggests that the attractive high *expected* wages (adjusted for the distribution of employment probabilities) in urban areas pull individuals into economic hubs, despite a long queue of workers already pursuing the same ideal. The net benefits of migrating to urban regions with higher potential wages (but also a greater risk of joblessness) outweigh a more secure, but lower wage in the subsistence rural sector.

On the labour demand side, firms in many cases do not relocate to regions where labour is cheap and relatively abundant (such as rural areas), but stay in highly productive regions. Economic hubs are therefore often centred on highly rewarded, productive inputs, rather than profiteering through finding locations with low input costs. Nominal wage differentials across regions in the United States are large and spatial patterns are persistent over the long-run; they correspond to similarly persistent patterns of labour productivity and innovation, which are in turn a function of agglomeration economies (Moretti, 2011). However, real wages (where local prices are accounted for, rather than national time variation) appear to have a more uniform spread, disincentivizing people to move to high income regions; the rigidity in this case is the high cost of living, particularly in terms of property.

The Rosen (1979) and Roback (1982) framework predicts that (in a scenario where labour is perfectly mobile) if a particular city experiences a productivity shock, this will reflect in both nominal wage and property price adjustments, so that real wages remain constant. Consequently, benefits of productivity shocks accrue in the form of capital rents, but none go to workers, and so real wages always equalize across regions. Moretti (2011) generalizes this model (through small elasticities of both labour and housing supply) to account for the case in which labour cannot move voluntarily. Such rigidities are relevant to economies such as South Africa, to the extent that apartheid regulations effectively limited housing and labour supply to blacks in regions that were designated for whites. Figure C1 in Appendix C shows that in this country long-run wages (that do not adjust for local prices) are spatially diverse. As predicted by the model, even real wages could, however, differ by region in long-run equilibrium. In particular, productivity shocks in regions where housing supply is less elastic than in other regions, can benefit property owners as opposed to workers. Apartheid-era urban

regions were productive but restricted, while homelands regions had fewer limitations on labour and residential supply to the black population group. Hence, this model predicts that in the pre-transition period, positive productivity shocks would accrue to land owners as opposed to workers. With the (relatively) free expansion of urban informal settlements in the post-apartheid period, this model supposes that shocks to productivity may benefit workers to a greater extent than was the case in the past.

By all expectations, South Africa should have experienced movements leading to a new spatial equilibrium that allowed workers to take advantage of freedom of movement. However, many spatial patterns from the past persist, with homelands remaining high unemployment regions (in contrast to both the Harris & Todaro (1970) and Moretti (2011) predictions). It is therefore likely that other rigidities (such as collective bargaining councils and barriers to entry into the informal sector) exist that raise the costs of job search in productive urban regions, so that even though internal migration has occurred, the post-apartheid equilibrium has not settled at a point of spatial uniformity.

Given the particular restrictions mentioned above, South Africa offers an interesting case study of the existence of separate local labour markets. Labour demand has traditionally been concentrated near mines and in metropolitan regions of white apartheid South Africa, while the dominant location of (unskilled) labour supply was purposefully decentralized to homeland regions by the apartheid regime. The former were small pseudo-independent "states" that were created by the apartheid government, where apparent black self-determination was instated. Black South Africans were effectively only citizens of these "states" (which were not recognized by the international community), even though the only viable labour market opportunities were available in the designated white areas of South Africa. A migrant labour system developed, whereby blacks were only allowed to work in white South Africa with permits, and were only recognized as temporary residents. Families stayed behind in the homelands, while only employed blacks could realistically migrate to industrial centres. However, migrant workers remain primarly connected to their sending households, and likely return home at selected intervals (Posel, 2010).

As a consequence, a spatially divided labour market resulted, with few viable opportunities in the homelands and unemployment rates that were approximately double those of other regions. As apartheid laws were repealed in the early 1990s (though free movement of workers from all race groups was allowed as early as 1986), it should be expected that the spatial duality of the labour market would diminish. Posel (2010) established that while patterns of temporary migration to economic centres continued well into the post-apartheid era, that a slow moving shift towards permanent settlement in the receiving (industrialized)

regions has started to emerge. Additionally, household income from remittances has started to decline, suggesting that a new spatial equilibrium could be emerging in post-apartheid South Africa. However, despite large movements of people to areas of greater economic activity, complete spatial equilibrium has not been achieved, with high unemployment still concentrated in rural areas. One potential explanation for this slow change is the important role that social grants have played in rural households' incomes. These grants have provided a public safety net that have allowed unemployment to persist in some regions, potentially crowding out the reliance on private remittance transfers (Klasen & Woolard, 2009).

Importantly, these factors inform our understanding of long-run spatial equilibria. Kingdon & Knight (2006a) conclude that homeland regions tend to have concurrently high levels of unemployment and relatively high wages, while a negative relationship exists in other parts of the country. Spatial separation has created distinct labour markets over the long-run. Will migration eliminate these differences over time? And is the negative relationship in non-homeland regions reflective of this long-run equilibrium or representative of short-run wage flexibility in the labour market. The rest of this chapter differentiates between these two effects.

2.2.2. Defining local labour markets

Local labour markets can potentially be defined as the geographic regions or demographic groupings from which workers source information and which they take into account in their wage setting decisions. In particular this includes the labour market conditions experienced by other peers within individuals' race groups and locality. This definition falls in line with the concept that Kingdon & Knight (2006b) used to test whether the non-searching employed were a part of the labour market. By way of wage curve estimates, they found that wage demands also factored in the stock of local discouraged workers, with large coefficients relating these variables (as opposed to only modelling the searching unemployed as surplus workers). Hence, they conclude, the non-searching are a part of the *de facto* labour market, even if official definitions do not acknowledge this.

The literature is, however, silent on what precisely defines "locality" of labour markets. If labour markets (in the wage curve tradition) are fairly homogenous over large regions, then wage setters would respond to unemployment rates in much the same way in large parts of the country. However, in countries where unemployment is highly dispersed, and where differences are large across regions (as well as time persistent), the local labour market should be narrowly defined, and accounting for fixed effects is important. As shown below, the unemployment is geographically dispersed in South Africa, so that the latter is an important consideration. In contrast, however, the more centralized bargaining is, the more homogenous

labour markets are likely to be across large regions, so that fewer regions of analysis could be used.

Often demarcations available in survey data dictate researchers' definition, and they also tend to assume that the smaller the region, the more "local" it is (Kingdon & Knight, 2006a). Similarly to Kingdon & Knight (2006b), it is possible to establish whether other regions or demographic groups should be included or excluded from a particular group of wage setters' "local" labour market, depending on whether their bargaining demands take those places' and groups' labour market conditions into account. The regional dimension takes on the form of an area of a certain geographic reach, for which wage setters take labour market conditions into calculation when bargaining: if the labour market is truly national in reach (and spatially undivided), then the national unemployment rate is of importance in influencing wage demands; if, however, it is of much smaller size (such as a district or city), then information from further afield is disregarded in wage setting decisions. Some studies note exactly this finding, by establishing that wages are more sensitive to national than sub-regional unemployment rates (Daouli et al., 2013). This is especially true where bargaining is centralized. In such contexts, bargaining extends the borders of what would otherwise be small local labour markets to include large areas, as wages are set across broader regions. Hence, slackness of the labour market in close proximity of workers does not influence wage setting decisions as much as they would had bargaining not been centralized, so that information from further afield is assimilated into what workers conceive of as a "local" labour market. This is confirmed theoretically and empricially in Nordic countries (Albaek et al., 2000), and has potentially important ramifications for the analysis of South Africa, given the large influence of bargaining arrangements.

The wage curve literature provides a tool to investigate this and other research questions in this dissertation, by considering how large regions must be for information on labour market conditions to flow and influence wage setting decisions. By delimiting units of measurement to "local" labour markets (which, importantly, have no uniform definition in these studies), the research that has followed on from the seminal work of Blanchflower and Oswald (1994), has consistently found that wages respond downwards to higher local unemployment rates.

Individual level (indexed by *i*) monthly earnings are regressed on regional unemployment rates as follows:

$$\log(earnings_{irt}) = \beta_0 + \beta_1 \log(local\ unempl\ rate\ _{rt}) + \gamma' x_{irt} + \mu_r + u_{irt} \quad (2.1)$$

The time index (*t*) indicates that individual level panel data could be used to also account for individual fixed effects; however, in many cases data limitations entail that estimates are

conducted with repeated cross sections, allowing researchers to control only for *regional* fixed effects, indexed by *r*. In cases where only single cross sections are available, fixed effects can only be introduced for regions with an area larger than those for which unemployment rates are calculated. As discussed below, this strategy may not remove all forms of unobserved spatial heterogeneity and lead to biased estimates.

Estimates of β_1 are most often around -0.1, reflecting the downward pressure on wages exerted by excess labour within a region. This typical elasticity suggests that as the unemployment rate increases by 10 per cent, wages fall by 1 per cent in response to labour market slackness. Because this figure is found in such a wide range of countries, it has been termed an "empirical law of economics" (Blanchflower & Oswald, 2008). Many studies therefore use it as a benchmark to evaluate whether wages are flexible or not. It can also, however, serve as an indicator of the definition of "local".

Labour markets may also be defined by actual daily flows of commuters (as opposed to temporary migration patterns across large distances). Individuals weigh up the potential wage benefits of distant jobs against the time and monetary costs of getting to these workplaces, optimising their decisions across space. Monetary costs include both daily transport expenditures and also the potential cost of relocating if individuals own homes far from their place of work. Simini et al (2012) adjust traditional gravity models, and show that commuting patterns can be accurately predicted based on population sizes between start and end locations, but importantly also the size of the population within a radius (centered at the starting location) equal to the distance between both locations. Each of these quantities reflect job availability in plausible alternative local labour markets in which individuals could conduct their job search at the same cost.

Recorded journeys to work have been used to delineate optimal "functional regions" for South Africa (Nel et al., 2008). By this definition, points located within delineated economic regions are more connected with other places that are also found within these confines, rather than with points outside. Because their data is related to commuting to work, the optimal regions they find by implication delineate local labour markets. Their analysis also shows that functional economic regions often do not correspond to current politically determined demarcations. The practical implication for researchers is that household surveys that are designed to measure regional labour market conditions do not necessarily capture local labour markets information correctly, as they are most often stratified by administrative rather than economically functional regions.

The empirical analysis in this study will test whether previous definitions of local labour markets that have been used by researchers are appropriate. Once these definitions are clarified, the ultimate objective is to establish whether the labour market is as flexible as the wage curve estimates of Kingdon & Knight (2006a) suggest.

2.3. Wage curves in high unemployment regions – the role of regional heterogeneity

In their recent entry in the Palgrave Dictionary of Economics on the wage curve as an "empirical law of economics", Blanchflower & Oswald (2008) express their surprise that the wage curve holds in a country with an unemployment rate as high as South Africa's. In fact, it is uncanny that the elasticity is just as high as cases measured in their original work on OECD countries (Blanchflower & Oswald, 1994). Their assertion is based on the evidence of Kingdon & Knight (2006a), who used the 1993 South African PSLSD data to estimate an elasticity of individual wages related to broad unemployment rates of survey clusters. Their estimates are also remarkably close to -0.1, despite South Africa's high unemployment rate.

In contrast, a comprehensive meta-analyses of the wage curve literature suggest that muted elasticities arise when unemployment is particularly high - though mostly non-linearities in unemployment are rejected statistically (Nijkamp & Poot, 2005). The intuition for this apparent contradiction derives from the poor potential for declining wages to clear particularly large labour surpluses (specifically if wages have to drop to below reservation or subsistence levels). Should unemployment be persistent and high in the long-run, it is possible that the wage flexibility implied by the wage curve does not conform to the "empirical law". Indeed, Fedderke's (2012) contribution starts off with the premise that "South African labor market conditions are unusual by international standards. High and persistent unemployment do not prevent real labor costs from rising." This emphasizes that not only do wages fail to *fall* in some high unemployment scenarios (as suggested by a non-linear wage curve), but in some circumstances they are even able to *rise.*

However, to fully reconcile the conflicting predictions on the nature of the relationship between wages and unemployment, it is necessary to separate short-run adjustments from long-run ones. It is apparent that in some settings, short-run non-responsiveness of wages to unemployment exists concurrently with long-run trade-offs, such as in the Nordic countries (Albaek et al., 2000). Without making provision for this in microeconometric models, the two effects are likely to be counfounded. The rest of this paper therefore attempts to understand how empirical wage curves should ideally be estimated to characterise various modern developing country labour markets, especially those with high levels of bargaining,

segmentation and unemployment. In particular, this paper turns to the role of (permanent) unobservable regional heterogeneity to account for long-run relationships. An additional object is to establish optimal definitions of regions and labour markets for which the empirical law holds.

Two reasons could explain why the existing evidence for South Africa conforms to international results from low unemployment countries, while it contradicts expectations for a high unemployment economy. Firstly, the work of Kingdon & Knight (2006a) documents a point in time before the large and steady rise in post-apartheid unemployment took place (Burger & von Fintel, 2014), and before many of the labour laws that they cite for apparent inflexibility were revised to become more strict. Secondly, because they are limited to work with a cross section, the effect they capture cannot distinguish between transient changes in the labour market and long-run factors. The result is that the estimates potentially confound these two effects and that (as in the case of Nordic countries), no short-run adjustments in wages exist, but that the negative relationship is an observance of a long-run equilibrium that manifests in the data. This is also the case in Greece where no transient wage flexibility is observed in relation to high unemployment, while a long-term trade-off exists (Daouli et al., 2013). The reasons in this case are similar, with high centralization of bargaining being linked to labour market inflexibility.

Kingdon & Knight (2006a) acknowledge potential difficulties by adding a quadratic in unemployment for the South African specification, so as to allow for a section on the unemployment-wage profile that is flatter than the rest of the curve. In most international contexts, significantly positive non-linearities are predicted only for sections of the wage curve that are not actually contained in data; the high South African unemployment rate, however, means that this potential phenomenon cannot be ignored.

Kingdon & Knight (2006a) further explore these issues by segmenting their models into high and low unemployment regions. Their analysis focusses on differences between previous “homeland” and “non-homeland” regions. An insignificantly positive wage curve relationship existed in homeland regions (with high unemployment and concurrently high wages after conditioning on relevant factors), while a strong negative wage curve holds in the rest of South Africa. Results such as these highlight the importance of regional heterogeneity where unemployment is not only high, but where it is widely dispersed along the spatial dimension. However, the latter finding still stands in contrast to assertions that South Africa has an inflexible labour market.

Magruder (2012) investigates another feature of the South African labour market that has a strong spatial definition, and which provides clues in the distinction between long-run and

short-run wage flexibility. Collective bargaining takes place in many industries, where dominant firms and unions set up wage agreements that often become applicable across entire regions. Under the provisions of the Labour Relations Act, the South African Minister of Labour has the discretion to extend a bargaining council agreement to an entire industry within a particular geographic jurisdiction, regardless of whether firms employ workers that were members of the unions that constitute the bargaining council. Consequently, districts in which wages are bargained up tend to witness lower employment creation, especially amongst smaller firms, which tend to be in disproportionately low concentration in South Africa.

Magruder (2012) uses these district-level spatial discontinuities to identify what can indirectly be viewed as behaviour that is contrary to a wage curve in post-apartheid South Africa, but through a specific channel: in the presence of a bargaining council, wages increase by 10-21% and employment drops by 8-13%. These wage premia results are very close to other estimates using alternative identification strategies, which show that bargaining councils are important for raising wages, as opposed to only firm-level union negotiations (Bhorat et al., 2012).

However, Magruder (2012) emphasizes that not all of unemployment in South Africa can be explained by a bargaining council effect, but acknowledges its long-term structural nature (Bhorat & Hodge, 1999). This distinction between the long and short-run is important for understanding the results of his study, and to inform the interpretation of wage curve estimates. While he finds an implied positive relationship between employment losses and wage increases in similar data[3] to that of Kingdon & Knight (2006a), it is likely that this represents only a transient wage curve relationship (through the collective bargaining channel), because of the appropriate treatment of time-invariant spatial fixed effects in his models.

The introduction of fixed effects, however, raises multiple questions, which have been addressed in the literature in various manners, often dictated by data availability. Firstly, what is the appropriate geographic definition of a "local" labour market, and is this the correct level at which to measure unemployment for wage curve purposes? Potential answers to this question have been discussed above. A related question is: which accompanying fixed effect should be used in a wage curve specification to account for unobserved spatial heterogeneity associated with labour market institutions (such as collective bargaining) and political

[3] He uses Labour Force Surveys, which were collected by Statistics South Africa in the first half of the 2000s decade. These shared a similar survey design with the 1993 PSLSD used by Kingdon & Knight (2006a). The essential difference is the luxury of the time element which Magruder (2012) enjoys in order to account for spatial heterogeneity effectively.

configurations (such as the former homelands system)? Secondly, what is the empirical importance of accounting for fixed effects, and which type of data is required to support such an analysis?

While the groundbreaking work on the wage curve (Blanchflower & Oswald, 1994) advocated the introduction of regional fixed effects in the standard specification, it did not have the same impact as established by the Scandinavian researchers (Albaek et al., 2000). In some cases the introduction of regional fixed effects amplifies the elasticity (see for instance Papps (2001)), while in others they are muted (see for instance Baltagi, Blien, & Wolf (2000)). Blanchflower & Oswald (1994: 181) emphasized their inclusion to distinguish between long-run and short-run differences in the wage-unemployment relationship.

Suppose that the population regression function contains regional fixed effects ($\mu_{r_j^*}$), and optimal local labour markets are defined by an unknown radius r^*as follows:

$$\log(earnings_{ir^*t}) = \beta_0 + \beta_1 \log(unemployment\ rate_{r^*t}) + \mu_{r^*} + u_{ir^*t} \quad \ldots(2.2)$$

Without accounting for regional fixed effects, estimates of the short-run wage curve coefficients deviate from the population coefficient, depending on the relationship between the unemployment rate and the fixed effects:

$$plim\hat{\beta}_1 = \beta_1 + \frac{Cov[\log(unemployment\ rate_{r^*t}); \mu_{r^*}]}{Var[\log(unemployment\ rate_{r^*t})]} \quad \ldots(2.3)$$

Given that $\beta_1 < 0$ and that the relationship between the unemployment rate and fixed effects is positive positive, $\hat{\beta}_1$ will be underestimated (or biased towards zero). Alternatively, if the long-run relationship (the correlation between unemployment and the regional fixed effects) is negative, $\hat{\beta}_1$ will be overestimated (or more negative than it should be).

A positive *long-run* relationship is empirically verified in the United States (along with a negative transitory wage curve) (Blanchflower & Oswald, 1994), while a negative long-run relationship persists in the Nordic states (with a statistically insignificant transitory wage curve) (Albaek, et al., 2000). A long-run negative relationship in the latter case explains why unemployment elasticities are overstated when not accounting for fixed effects (though small numbers of regions in their analysis could be confounding this analysis).

Why, then do the Nordic results differ from the "empirical law", and are there lessons to be learnt for South Africa? Albaek, et al. (2000) emphasize the role of centralized bargaining in driving these results. Under highly centralized bargaining regimes, wages are likely to be unresponsive to higher *local* unemployment, as they are set at a more central regional level.

Recent firm-level evidence for Germany estimates separate wage curves for workers who negotiate at the firm level and workers that are covered by sector-wide bargaining agreements (Blien, et al., 2013). Wages are not at all responsive to local unemployment when workers benefit from sectoral agreements; however, unionized workers that negotiate only at the firm level do display wage curve behaviour. The implication is that the further removed the negotiation is from the firm, wages tend to become less flexible, as agreements do not take local labour market conditions into sufficient account.

Given the provision for sectoral determination in South Africa, the high wage flexibility found by Kingdon & Knight (2006a) is surprising, and requires further investigation. It is likely that spatial heterogeneity has not been effectively accounted for. Some sectoral bargaining councils are centralized at the national level in South Africa, while others ecompass large districts. In each of these cases, bargaining is centralized to a level beyond the immediate firm or town, so that wages may also not be as responsive to "local" labour market conditions. The following section will follow up with a review of the literature, with the role of bargaining in mind in generating results different from the "empirical law".

2.3.1. A meta-analysis of wage curve studies and regional heterogeneity

Nijkamp & Poot's (2005) existing meta-analysis shows that including regional fixed effects in wage curve specifications reduces elasticities by an average of 0.035 percentage points, though this fall is not statistically significant across the specifications they survey. Accounting for long-run wage setting dynamics therefore does not detract from the wage curve as a transitory empirical law in most instances. However, most of the studies they survey are taken from developed country evidence, where unemployment is not as high, and also not as dispersed across regions. They also do not explicitly explore the dimension of centralized bargaining in determining different long-run equilibria, as noted above. However, one of their examples specifically emphasizes the role of centralized bargaining in distinguishing between a long-run and a short-run wage curve (Albaek et al., 2000). The object of this section is to highlight the main features of their meta-analysis, but to also study whether their conclusions on fixed effects can be potentially differentiated along the lines of the degree of centralized bargaining within an economy. Additionally, the number of "local" labour markets in each study is highlighted, in order to understand whether researchers use small or large regions to find their effects, and by implication to learn about the size of local labour markets in wage curve studies.

Most studies utilise repeated cross sections, and present evidence of wage curves with and without regional fixed effects; in these studies it is possible to include fixed effects at the

same level at which unemployment rates are calculated due to the time dimension that is present in the series of data. In a few cases a single cross section survey is used by necessity, and fixed effects are included for regions that are geographically larger than the level at which unemployment rates are calculated. This raises the question whether fixed effects for regions larger than those for which unemployment rates are estimated can effectively remove the long-run component of the wage setting relationship. Before exploring these dimensions in light of Nijkamp & Poot's (2005) overview, the context of South Africa is sketched.

Kingdon & Knight's (2006a) work is a case in point, where fixed effects potentially did not separate long-run from short-run wage curve effects. In the absence of multiple datasets with geographic variation, they resort to using one cross sectional survey. Consequently, they are not able to introduce regional fixed effects at the same level of aggregation as the unemployment rates they measure. As a second best option, they include fixed effects for regions with larger geographic definition[4]. However, even within these larger regions, substantial heterogeneity exists in unemployment rates. Despite some provinces clearly having higher unemployment rates on average, it is evident that the heterogeneity within these geographic units is non-negligible. Notably, unemployment is highly correlated with the apartheid homeland demarcations.

Figure 2.1 shows that this is true even in the 2000-2004 period, some years after the homeland system was abandoned by the post-transition regime. Many provinces (as in Figure 2.3), display substantial variation in unemployment *within* their boundaries, with homeland sub-regions standing out as high unemployment areas. This contrast is particularly stark in the northernmost Limpopo province of South Africa. While this province is a high unemployment region as a totality (Figure 2.3), it also contains many magisterial districts with substantially lower levels of unemployment that are directly adjacent to high unemployment districts. As discussed above, this persistent situation is potentially the product of social grants that have expanded more rapidly in homeland regions (Pienaar & von Fintel, 2014); unemployed individuals live in households with recipients of grants and remain isolated from main labour market centres (Klasen & Woolard, 2009). Alternatively, the costs of migration may be high, as existing land rights would be forfeited if individuals decided to move from the homelands to enter the urban labour market.

Not accounting for this level of heterogeneity is potentially important for wage curve analyses. It is firstly evident that a province potentially cannot define a "local labour market" purely by the differences in unemployment within these administrative regions. However, if

[4] They include provincial fixed effects (the demarcations are shown in Figure 2.3), while estimating unemployment rates for survey clusters.

wages setters disregard these differences and factor average unemployment rates within geographically dispersed social networks into wage bargaining, then a province or even a larger region could nevertheless be regarded as a local labour market. The alternatives should be investigated by conducting various wage curve analyses, testing the robustness of estimates to the labour market demarcation chosen.

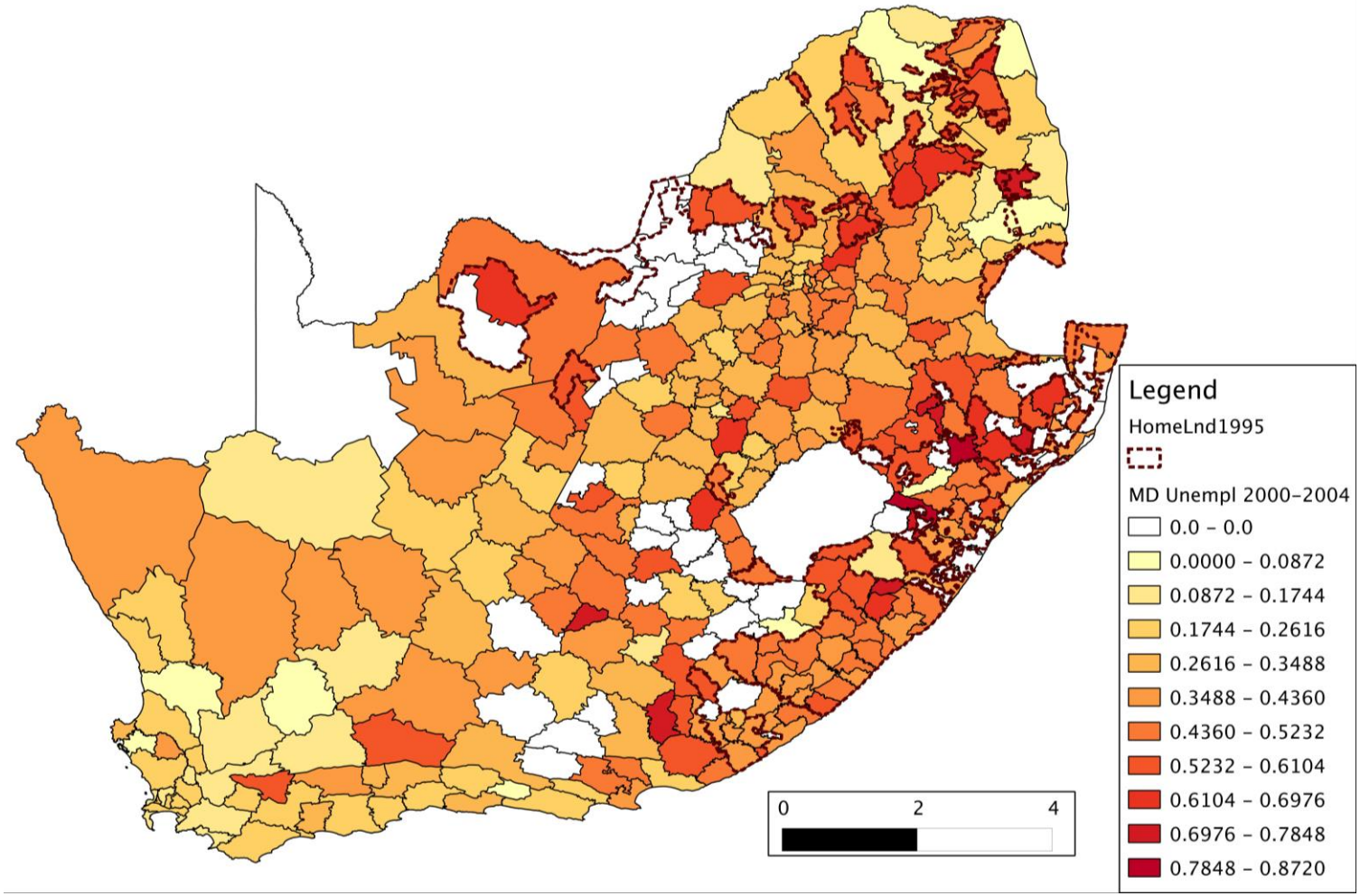

Figure 2.1 Broad magisterial district unemployment rates - 2000-2004

Source: Own calculations from LFS2000-2004. The unemployment rate is for the entire labour force within the region, calculated by the broad definition and pooled over the period.

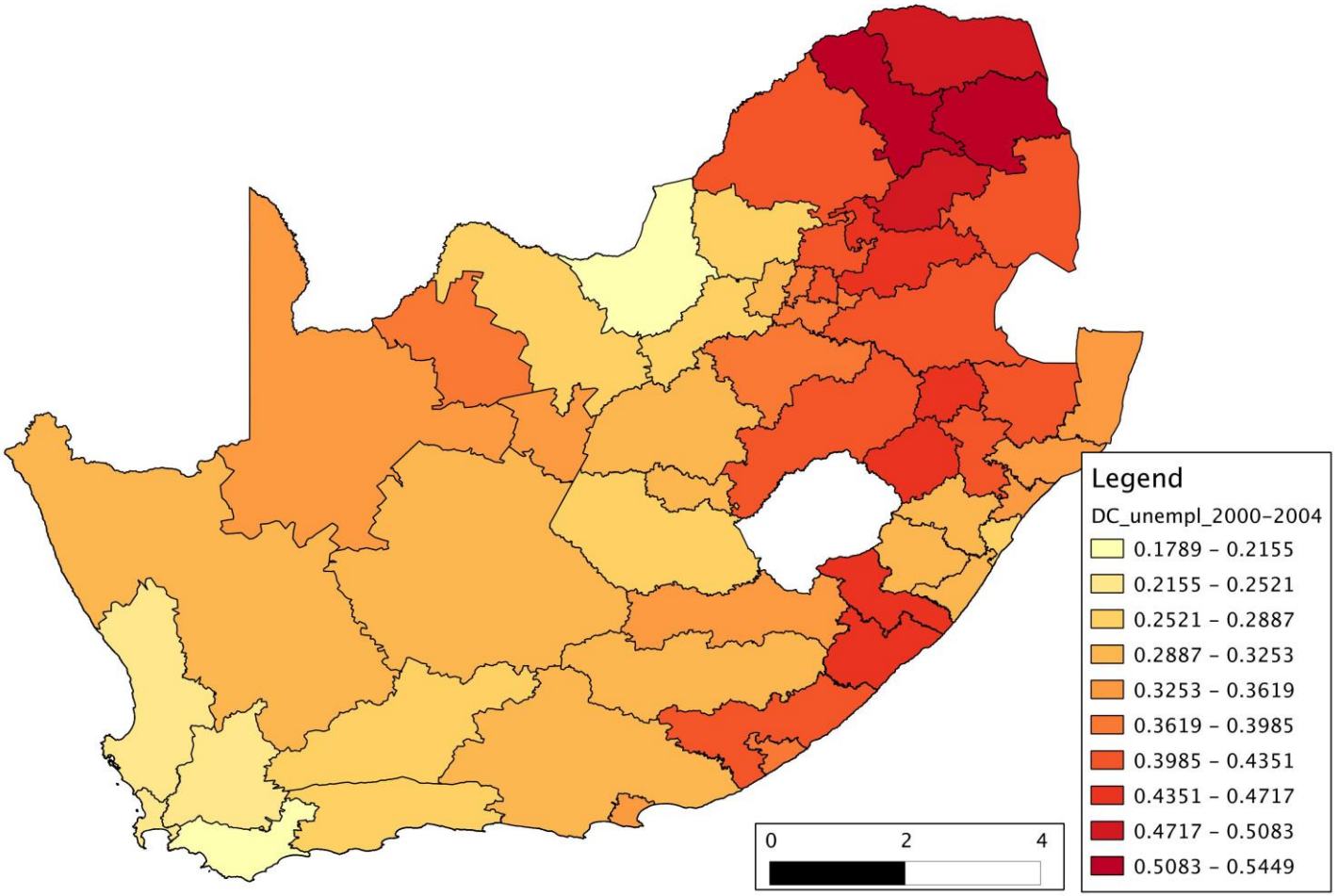

Figure 2.2 Broad district council unemployment rates - 2000-2004

Source: Own calculations from LFS2000-2004. The unemployment rate is for the entire labour force within the region, calculated by the broad definition and pooled over the period.

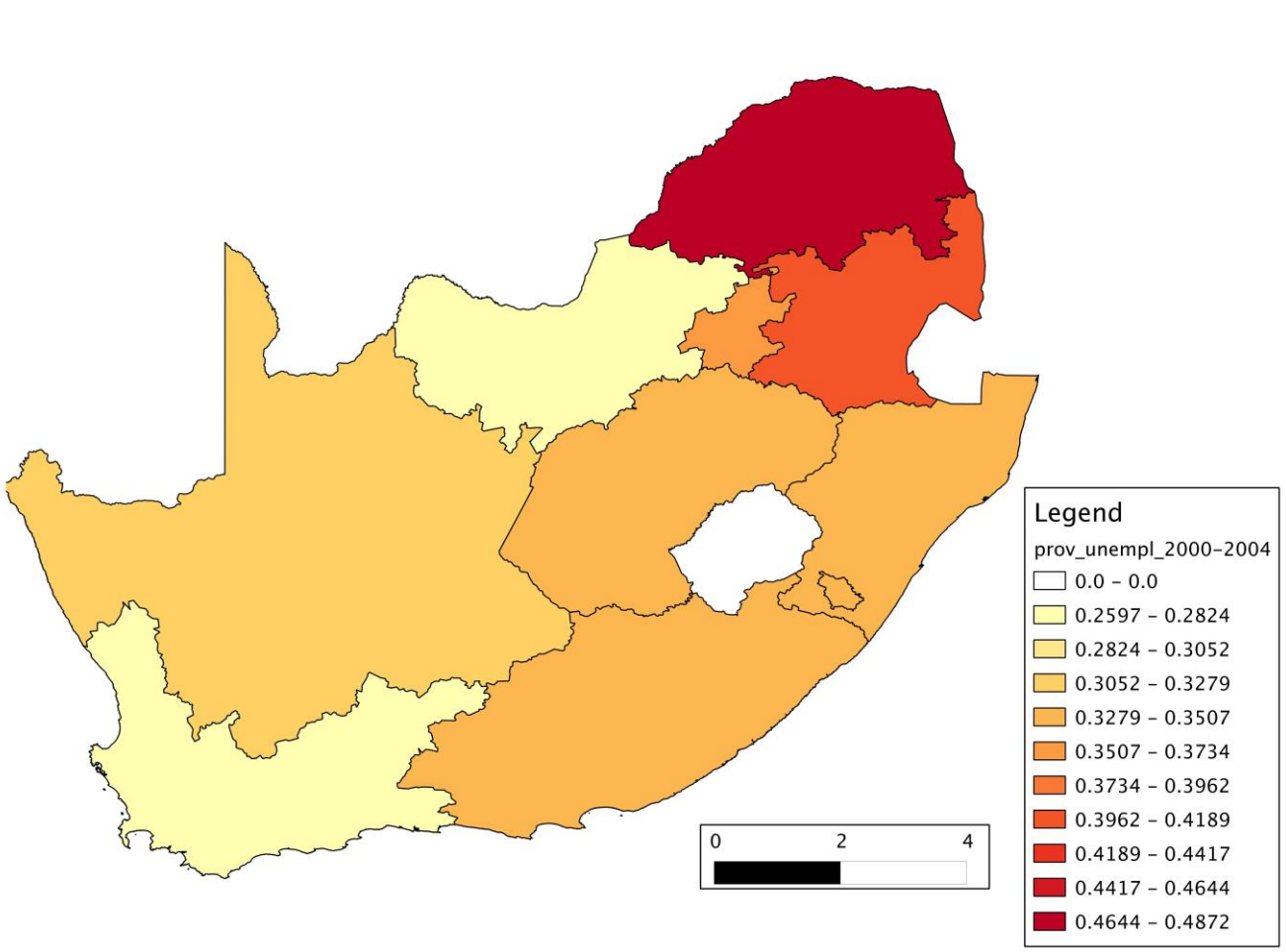

Figure 2.3 Broad provincial unemployment rates - 2000-2004

Source: Own calculations from LFS2000-2004. The unemployment rate is for the entire labour force within the region, calculated by the broad definition and pooled over the period.

Furthermore, if it is true that labour markets are defined by areas smaller than a province, the heterogeneity *within* provinces will not be accounted for by introducing fixed effects representing this larger geographic level. Given that unemployment is of a long-term structural nature in South Africa, it is not sure that incorrectly defined fixed effects can purge the wage curve relationship from this feature adequately.

Finally, Card (1995) highlights issues with degrees of freedom when regions are few: to illustrate, should 9 provincial unemployment rates be used to model individual wages, the wage curve elasticity is not based on the number of individuals, but 9 multiplied by the number of surveys used. As a result, standard inference using aggregated data to explain individual level outcomes is potentially misleading, if the number of aggregate units is small (Moulton, 1990). Using smaller regions would have the benefit of more units of observation to cover the map, and which reduces problems for statistical inference. In this study, standard errors are clustered at the relevant geographic unit to account for this problem, as is done elsewhere in the wage curve literature (Sanz-de-Galdeano & Turunen, 2006)[5].

Kingdon & Knight (2006a) take account of the first and third concerns in their study by estimating unemployment rates at the survey cluster level, which covers substantially smaller areas than provinces. However, their use of cross sectional data does not allow them to address the second concern, so that it is likely that sufficient long-term heterogeneity remains unabsorbed by fixed effects estimation. Other studies suffer the same limitations (Blanchflower & Oswald, 1990; Winter-Ebner, 1996). In each case, the introduction of fixed effects at a level of aggregation larger than the unemployment rate reduces wage curve elasticities, though does not render them insignificant (as is the case in other studies). However, the potential remains that the elasticity remains overstated. This possibility is discussed in more detail below.

While Nijkamp & Poot (2005) conclude that the introduction of fixed effects does not have an overall impact on the elasticities that they survey, some specific differences nevertheless emerge. As noted above, their overview does not pay much attention to the problem of using few (but large) regions to proxy for labour markets, and also does not consider the role that fixed effects can play when centralized bargaining is high. The rest of this section provides some re-interpretation of their evidence, which also potentially explains why Kingdon & Knight (2006a) obtain results that are not expected for an economy such as South Africa. Table 2.1 draws together most of the estimates they review (bar for those from the book of Blanchflower & Oswald (1994)), but now explicitly focusses on study dimensons relevant to

[5] Unemployment rates should ideally also be treated as generated regressors, with more conservative standard errors being implemented. While acknowledging this shortcoming, it is beyond the scope of this study to consider the effects of this potential problem on statistical inference.

the current question, namely the number of regional unemployment rates, the number of regional fixed effects, and also the potential for centralized bargaining to dampen short-run wage fluctuations. The latter is achieved by comparing countries by measures of the degree of centralization of their collective bargaining systems (Iverson, 1998) and the prevalence of central bargaining (Driffill, 2006). The purpose of this attempt is to highlight under which circumstances various empirical strategies generate wage curves that should be scrutinised more carefully, and to give potential explanations for some of the deviations from the norm.

Starting with studies of countries that have lower levels of centralized bargaining, US and UK estimates are robustly in line with the wage curve, even when accounting for regional fixed effects. The exception is the repeated cross section estimate of Blanchflower & Oswald (1990), which becomes statistically insignificant when fixed effects are introduced. However, this exposes a problem of limited degrees of freedom rather than bias (Card, 1995), as only 11 regional unemployment rates were constructed. This questions the validity of statistical inference rather than economic factors that lead to the apparent inflexibility that is measured. The specifications of Wagner (1994) suffer the same problem, so that the insignificance after fixed effects are introduced can potentially not be attributed to the relatively high centralization of bargaining that prevails in Germany.

However, other estimates for (East and West) Germany are also rendered insignificant by *regional* fixed effects (Baltagi, Blien, & Wolf, 2000; Baltagi & Blien, 1998), despite defining larger numbers of regions. As a result, the potential role for a centralized bargaining channel re-emerges. Other estimates for Germany (Pannenberg & Schwarze, 1998; Buettner, 1999) remain significant but small after conditioning on regional dummies. For Germany (with relatively highly centralized bargaining institutions), the evidence could be consistent with the claims of Albaek, et al. (2000) that wage curve elasticities are either small or insignificant in the presence of strongly centralized bargaining.

However, antipodean studies (in regions with only slightly lower degrees of centralized bargaining than Germany) do not follow this pattern. Australian repeated cross sectional data (Kennedy & Borland, 2000) yields smaller (yet significant) estimates of the elasticity after introducing fixed effects, despite a small number of defined regions and moderate levels of centralized bargaining. In the case of New Zealand, the elasticity becomes more strongly negative (Papps, 2001). Belgian estimates (Janssens & Konings, 1998) follow a similar pattern to those from Australia, with similar limitations in terms of numbers of regions and within a similar bargaining environment.

Turning to highly centralized bargaining regions, Albaek, et al. (2000) attribute their insignificant elasticities (after introducing regional fixed effects) to the centralized bargaining

systems that operate in the Nordic countries. However, small degrees of freedom could just as likely be playing a role in this case, other than the reasons that they offer. Their choice of few but large regions may, however, be justified on the grounds that "local labour markets" could be potentially larger where central bargaining prevails. Furthermore, Austrian estimates (Winter-Ebner, 1996) – with a sufficient number of regions (though a small number of fixed effects) – remain statistically significant once controlling for region. This is true despite it being at the extreme of the centralized bargaining spectrum; however, the lack of additional cross sections entails that higher level regional fixed effects were not likely to mop up the full degree of heterogeneity that was present in the data. This is similar to the case of South Africa.

Overall, then, evidence for centralized bargaining as the mechanism for insignificant or small short-run estimates does carry some credibility, and is potentially masked by the existing meta-analysis' inclusion of many studies with low to moderate degrees of centralized bargaining, and in some instances low effective statistical degrees of freedom (Nijkamp & Poot, 2005). The evidence on the number of local labour markets and fixed effects is also not unambiguous from this meta-analysis. However, understanding levels of heterogeneity within these regions (perhaps also along the lines of bargaining councils, *inter alia*) is a potential route to uncovering why definitions of regions are as important for wage curve analysis as their presence in the literature suggests. This paper will not repeat a meta-analysis taking these specific factors into account, but will attempt to illustrate these features using a series of repeated South African cross sections. In this context, collective bargaining plays an important role in the wage setting mechanism (Bhorat et al., 2012; Magruder, 2012). As a result, short-run wage flexibility could be limited, while a long-run trade-off may nevertheless emerge. Spatial heterogeneity is therefore a potentially important component of the analysis.

Table 2.1 Summary of wage curve studies and their regional fixed effects

							Centralization		Collective bargaining coverage*		
Author	**Country**	**Period**	**Data type**	**Number of u% rates**	**FE Regions**	**Impact of introducing regional FE**	**Index§**	**Score***	**1980**	**1990**	**2000**
Blanchflower & Oswald (1990)	Britain	1981	CS	65	9	Elasticity remains significantly negative	0.177	3+	70	40	30
Blanchflower & Oswald (1990)	Britain	1983-1987	Rep CS	11	11	Elasticity becomes insignificant	0.177	3+	70	40	30
Wagner (1994)	W Germany	1979;1985	Rep CS	10	10	Elasticity becomes insignificant	0.377	5--	80	80	90
Wagner (1994)	W Germany	1984-1990	Rep CS	9	9	Elasticity becomes insignificant	0.377	5--	80	80	90
Bratsberg & Turunen (1996)	USA	1979-1993	Panel	1376	1376	Elasticity becomes smaller, but remains significant	0.071	2	26	18	14
Winter-Ebmer (1996)	Austria	1983	CS	99 regions x 31occupations	9	Elasticity becomes smaller, but remains significant	0.431	6	95	95	95
Partridge & Rickman (1997)	USA	1972-1991	State panel	48	48	Elasticity is positive, but becomes smaller	0.071	2	26	18	14
Baltagi & Blien (1998)	Germany	1981-1990	Panel	142	142	Not significant for all workers, except when instrument added	0.377	5--	80	80	90
Janssens & Konings (1998)	Belgium	1985-1992	Indiv Panel	11	11	Remains negatively significiant	0.321	4	90	90	90
Pannenberg & Schwarze (1998)	E Germany	1992-1994	Indiv Panel	35	35	Elasticity negatively significant with regional FE, but not with individual FE					
Buettner (1999)	W Germany	1992	Reg Panel	325	325	Small but negatively significant	0.377	5--	80	80	90
Albaek et al (2000)	Nordic Countries	1989-1993	Rep CS	8 to 19	8 to 19	Large region u% insignificant with same FE; limited effect for district level	0.459 to 0.538	5- to 5	70 to 90	70 to 90	70 to 90
Baltagi et al (2000)	E Germany	1993-1998	Rep CS	114	114	Fixed effects knock significance, instruments bring it back					
Kennedy & Borland (2000)	Australia	1982-1995	Rep CS	7	7	Elasticity becomes smaller, but remains significant		4	80	80	80
Papps (2001)	N Zealand	1986-1996	Rep CS	60	30	Elasticity becomes stronger		4	60	60	25
Blanchflower (2001)	Eastern Europe	1991-1997	Rep CS	6 to 42	6 to 42	Elasiticty becomes smaller: significant in most estimates; sometimes insignificant					

Source: Own summary of studies included in meta-analysis of Nijkamp & Poot (2005). CS = Cross Section. Centralization index from §Iverson (1998); centralization score & collective bargaining coverage indicators from *Driffill (2006).

2.3.2. Regional heterogeneity and wage curve bias

This section proceeds to outline how the wage curve specification can be distorted by inappropriately accounting for spatial heterogeneity, and furthermore considering the impact of measuring unemployment at the incorrect spatial dimension. In each case, the potential bias for estimates is considered in order to establish the most credible way to estimate the relationship using South African data.

As before, suppose that the true wage curve is represented by:

$$\log(earnings_{ir^*t}) = \beta_0 + \beta_1 \log(unempl\ rate_{r^*t}) + \gamma' x_{ir^*t} + e_{ir^*t} \quad \ldots(2.4)$$

where $e_{ir^*t} = \mu_{r^*} + \lambda_t + u_{ir^*t}$, i and t index individuals and time respectively, and r^* indexes a set of geographic reference regions. The latter are true (but unknown) functional local labour markets, for which individuals take the full set of information into account in their wage demands. The size and number of these regions within a country depends on how spatially polarized the labour market is, but may also be influenced by social network connections that bridge regions with very different labour market conditions. Should labour market circumstances be fairly homogenous across space, and workers able to migrate freely into another region or industry, the optimal region may be fairly large. In the South African case, these optimal regions have the potential for being smaller than in other countries, owing to the spatial segregation introduced by the homeland system. While labour market movement between these regions has been liberalised after apartheid, the spatial patterns along very small regional lines still persist (as is evident in Figure 2.1). However, even collective bargaining and temporary migration (Posel, 2010) can extend the labour market to encompass regions that are spatially far apart. This could occur through the working of social networks, which are found to be very important for job search and hiring in South Africa (Hofmeyr, 2010; Schoer, et al., 2014).

Should equation (2.4) be estimated without adding any regional fixed effects, the standard result holds that coefficients of interest may be biased and inconsistent if any of the covariates are correlated with the unobserved heterogeneity that is absorbed into the error term (Wooldridge, 2010). As highlighted by equation (2.3), bias depends on the long-run relationship between wages and unemployment.

2.3.2.1. Fixed effects for non-optimal regions

Suppose, however, that a lacking time dimension in the data disallows fixed effects to be added at the same level of spatial aggregation at which the unemployment rate is calculated. Where spatial differences are important, it may be preferable to account for heterogeneity at a

geographic level that is typically larger than the optimal labour market, a strategy followed by Kingdon & Knight (2006a). Suppose that it is still possible to calculate unemployment rates for true local labour markets, but that fixed effects for larger regions and industries are included as follows:

$$\log(earnings_{ir^*t}) = \tilde{\beta}_0 + \tilde{\beta}_1 \log(unemployment_{r^*t}) + \tilde{\gamma}' x_{ir^*t} + \tilde{e}_{irt} \quad \ldots (2.5)$$

where $\tilde{e}_{irt} = \mu_r + \tilde{\lambda}_t + \tilde{u}_{irt}$ and $\mu_r = \frac{1}{n_{r^* \subset r}} \sum_{r^* \subset r} \alpha_{r^*} \mu_{r^*}$ is now a weighted "average" of the fixed effects of all of the true unknown sub-regions (r^*) of which the larger region (r) is comprised, with the optimal region specific deviations (ξ_{r^*}) from this average being absorbed into $\tilde{u}_{irt} = u_{ir^*t} + \xi_{r^*}$.

Hence,

$$\mu_{r^*} = \mu_r + \xi_{r^*}$$

so that

$$plim\widetilde{\beta_1} = \beta_1 + \frac{Cov[\log(unemployment_{r^*t}); \xi_{r^*}]}{Var[\log(unemployment_{r^*t})]} \quad \ldots (2.6)$$

where

$$Cov[\log(unemployment_{r^*t}); \xi_{r^*}]$$

$$= Cov[\log(unemp_{r^*t}); \mu_{r^*}] - \frac{1}{n_{r^* \subset r_j}} \sum_{r^* \subset r_j} \alpha_{r^*} \, Cov[\log(unemp_{r^*t}); \mu_{r^*}] \quad \ldots (2.7)$$

Given that this sub-optimal fixed effect does, not "mop up" all of the unobserved heterogeneity in a true but unobserved labour market, bias emerges, as in (2.6). The first term of equation (2.7) determines the inconsistency that would arise if no fixed effects were included in the specification, as highlighted in (2.3). The second term in (2.7) represents the potential reduction in this bias that results from at least including the rougher fixed effects. However, suppose that large degrees of spatial heterogeneity exists within the larger district (r) than in each of its component districts (r^*), so that one district (say) with a large population (large α_{r^*}) is a high unemployment-high wage district (in the Harris-Todaro (1970) sense), while other districts have (say) close to zero correlations between unemployment and the long-run wage fixed effect. The consequence of this would be a large between district correlation (large first term), and a small sum of within district correlations (small second term), so that the overall correlation between unemployment and the unobservable heterogeneity is not eliminated by the rougher fixed effect. Essentially this suggests that if area fixed effects attempt to account for large levels of heterogeneity across

sub-regions, it stays likely that estimates will remain biased and inconsistent if this heterogeneity *within* the larger regions is substantial. Should μ_r account for much of the variation in μ_{r^*} as opposed to ξ_{r^*}, this strategy will result in minimal bias.

2.3.2.2. Optimal Labour Market Size for Wage Curve Analysis

Since most studies in the wage curve literature are silent on what should be the size of the "optimal" labour market, and no clear prescriptions exist at which level unemployment rates should be calculated, other potential (unnoticed) problems may arise as a result. Wagner (1994) and Kingdon & Knight (2006a), for instance, recommend that unemployment rates should be measured at lower levels of aggregation to reflect "local" labour markets. This, however, assumes that other factors such as centralized wage setting and demographic networks across large distances do not galvanize small regions into larger definitions of "locality". Presumably, most applied work relies on the information of regions that is available in survey data, which is often dictated by the sampling design, or political and administrative divisions. Kingdon & Knight (2006a) state this explicitly, noting that their choice of locality was data-driven and not based on institutional knowledge. Where local labour markets naturally fill these boundaries (such as where minimum wages and bargaining are defined regionally), this is unproblematic.

However, extending the analysis above, it may be that unemployment rates that are calculated for larger than optimal geographic regions, could compromise the estimation of the elasticity if labour markets are truly small[6]. Suppose now that unemployment is calculated for a larger region (r) than is optimal and used to estimate (2.8) below. Unemployment is now assumed to be measured with error ($\omega_{(r^*\subset r)t}$) that is multiplicative with the unemployment rate of the unknown optimal labour market region (r^*):

$$\log(earnings_{irt}) = \hat{\beta}_0 + \hat{\beta}_1 \log(unemployment_{rt}) + \hat{\gamma}' x_{irt} + \hat{e}_{irt}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \log\left(unemployment_{(r^*\subset r)t}\omega_{(r^*\subset r)t}\right) + \hat{\gamma}' x_{irt} + \hat{e}_{irt}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \log\left(unemployment_{(r^*\subset r)t}\right) + \hat{\gamma}' x_{irt} + \hat{v}_{irt} \quad (2.8)$$

$$\text{where } \hat{v}_{irt} = \frac{1}{n_{(r^*\subset r)}} \sum_{(r^*\subset r)} \alpha_{r^*}\mu_{r^*} + \xi_{r^*} + \lambda_t + \hat{\beta}_1 \log(\omega_{(r^*\subset r)t}) + u_{irt}$$

so that

[6] While this discussion focuses on regions that are larger than optimal, the results transfer to a situation where regions may be smaller than optimal, which also induces measurement error.

$$plim\hat{\beta}_1 = \beta_1 \frac{Var[\log(unempl_{rt})]}{Var[\log(unempl_{rt})]+Var[log(\omega_{(r^*\subset r)t})]} + \frac{Cov[\log(unempl_{r^*t});\xi_{r^*}]}{Var[\log(unempl_{r^*t})]} \quad (2.9)$$

This specification now yields a composite error term with two additional components that can be potentially correlated with unemployment. The second term in equation (2.9) is associated with the incorrect specification of fixed effects, while the first is of interest with regards to sub-optimal labour market size. Under classical measurement error assumptions, the inappropriate size of the labour market will reduce the absolute value of wage curve estimates. This will be true when labour markets are defined as either too small *or* too large, compared to the optimal size. As $Var[log(\omega_{(r^*\subset r)t})] \longrightarrow 0$, or as measurement error (due to incorrect demarcations) becomes equally dispersed across space, the bias is removed. As a result, the maximum wage curve estimates (in absolute value) will be obtained for labour markets that are most appropriately defined in the data. Any demarcation smaller or larger in area will yield a muted wage curve elasticity. For definitional purposes, therefore, locality can be defined as regions for which wage curve elasticities are largest.

Each of these measurement concerns can be addressed by respectively introducing appropriate fixed effects in order to isolate the transitory wage curve effect, and by defining locality according to geographic regions that are known to operate as separate labour markets. However, knowledge about optimal labour market size is not necessarily available, as acknowledged by Kingdon & Knight (2006a). Alternatively, one could instrument to account for this type of measurement error. Apart from inappropriate reach, choosing local labour market demarcations that are too small may also lead to attenuation bias related to the sampling frame: if surveys are stratified to be representative of geographic regions that are larger than chosen areas, unemployment rates may be constructed with too few observations to reflect the population of that place. A trade-off exists in calculating representative unemployment rates for fewer, larger regions, and having greater cross sectional variation achieved by calculating more unemployment rates for smaller regions. Instrumentation is therefore vital to account for all types of measurement error.

While instrumental variables (IV) are commonly used in the wage curve literature, the purpose is never explicitly related to measurement error. Rather, researchers instrument to account for possible bi-directional causality that introduces additional bias to estimates. Kingdon & Knight (2006a) argue that wages that are bargained beyond market-clearing levels induce higher unemployment, a proposition that others have also suggested to be true for South Africa (Fedderke, 2012). Consequently, estimates of wave curve elasticities will be biased toward zero. Kingdon & Knight (2006a) illustrate this by their use of cluster-level indicators as instruments.

The standard IV approach followed in this literature assumes that local unemployment rates are predetermined, so that time lags thereof are used as instruments for current values (Baltagi et al., 2012). In the absence of a time dimension in the data, Kingdon & Knight (2006a) did not follow this route. While fixed effects reduced wage curves estimates to insignificance in Turkey, adding lagged instruments brings them back in line with the normal benchmark of -0.1 (Baltagi et al., 2012). This confirms the direction of bias resulting from reverse causality. However, IV results ignore the measurement error issues discussed here. If, for instance, the same incorrect labour market demarcation is used in *all* periods, measurement errors will surely be correlated over time, rendering the IV endogenous and estimates inconsistent. Additionally, if a large component of unemployment rates is time persistent, adding fixed effects will strip substantial variation from both the current and lag periods. The results in a weak instrument. Each of these factors calls for alternative instrumentation strategies. This paper explores the use of two other instruments[7].

Firstly, a spatial discontinuity is exploited: unemployment rates jump at former apartheid homeland borders. They increase, the further they are from the border *within* homelands, while they decrease with distance *outside* the homelands. However, this instrument is time-invariant and correlated with labour market demarcations, so that it is not compatible with fixed effects estimation. Secondly, spatial lags of unemployment rates are used as instruments. Magisterial district unemployment rates are weighted within a pre-determined radius (of 200km) from the centroid of the district[8]. Weights are represented by the inverse of the distance between the centroid of one magisterial district from the centroid of another. In so doing, information that is spatially distant is given less weight in the composite unemployment rates, under the assumption that variables have greater influence on each other, the closer the geographic proximity (Arbia, 2006). Magruder (2012) instruments for bargaining council status in a magisterial district with the status in a larger district council, similarly to this approach. However, if the local labour market is in actual fact larger than the boundaries of the magisterial district, wages will be directly influenced by unemployment rates further afield, so that this may potentially result in an endogenous instrument. These eventualities are discussed where necessary.

[7] A third approach was also followed unsuccessfully, and is not shown in the text. Rainfall deviations from a long-run trend, as collected by Matsuura & Willmot (2012) are potentially correlated with labour market conditions in developing countries. However, given the diminishing role of agriculture in the economy, it is no surprise that this was a weak instrument, regardless of specification.

[8] Robustness checks use other radii. However, the instrument is weak at other intervals.

2.4. Data and approach

To explore these potential problems with wage curve estimates, we turn to the context of Kingdon & Knight (2006a), who find a wage curve for South Africa using cross sectional data in 1993. While they are able to estimate unemployment rates for fairly small geographical regions – survey clusters, which are smaller than magisterial districts (as in Figure 2.1) – their use of a cross section survey prevents them from controlling for fixed effects at that same spatial level. They therefore estimate approximately 360 cluster level unemployment rates and include only 9 provincial fixed effects.

This study employs a series of cross sections from South Africa's Labour Force Surveys, enumerated twice yearly in March and September by Statistics South Africa. The period spans September 2000 to March 2004, and is able to capture short-run changes in earnings and unemployment. These particular rounds have been chosen because they allow for the identification of individuals' place of residence by various classifications (the smallest being magisterial districts as in Figure 2.1, then larger district councils as in Figure 2.2 and finally provinces as in Figure 2.3[9]). The various demarcations allow for potential local labour markets of various sizes to be differentiated, as well as multiple levels of regional heterogeneity. Furthermore, this period is known for consistent measurement of labour market status and earnings over time (Burger & Yu, 2006). While this time was marked by robust economic growth, unemployment continued to rise concurrently. It is therefore of interest to know whether wages moderated in response to growing unemployment (or whether they tracked positive economic growth more closely). Furthermore, the empirical analysis seeks to understand whether existing high long-run unemployment influences the wage setting relationship.

Unemployment rates are based on the broad definition, following the recommendation of Kingdon & Knight (2006b). Discouraged workers are classified as part of the labour market through their similarities to the searching unemployed and also because wage setters take the stock of discouraged workers into account when framing their decisions. Individual monthly log earnings are used with a log-log specification, so that the coefficient of interest represents

[9] Magisterial districts are administrative boundaries that were adhered to in the apartheid era. However, many post-apartheid sampling frames used these small units, and data in various years can be linked at this geographic level. District councils are new administrative boundaries that were introduced in the post-apartheid era for the purposes of local government: they oversee smaller local municipalities. These 55 regions are larger than the magisterial districts, and are mostly comprised of a number of former magisterial districts. South Africa has a second tier of governance, divided into 9 provinces since 1994. Other rounds of the LFS cannot be linked to magisterial districts or other small regions, without additional information. Later datasets, such as the National Income Dynamics Study, enable precise location of households if secure data is retrieved. At the time of writing, this option was not yet available.

an elasticity. Standard Mincerian controls are included, as well as region-level covariates, in much the same way that Kingdon & Knight (2006a) did.

Initially, specifications probe the question whether wage setters are more likely to respond to unemployment rates of small regions within which they live (magisterial districts), or whether they take information from larger areas into account. Should the latter be true, it suggests that labour markets are not as small as magisterial districts, and that mechanisms that centralize wage setting (such as bargaining and networks) are at play. In each of these specifications, the role of spatial heterogeneity is tested, to understand whether the high levels of wage flexibility previously recorded for South Africa is transitory or long-term. Instrumental variables account for using potentially inappropriate labour market demarcations and other measurement errors, as well as reverse causality. Results are differentiated by race and gender, in order to establish whether demographic groups have specific responses to (geographically) local labour markets in their wage demands[10]. If labour markets are still segmented by race (as was the case under the apartheid dispensation), then it can be expected that blacks respond to overall labour market conditions most acutely, as they also make up the majority of the population (both the employed and unemployed). Other groups, in contrast, may not factor overall labour market conditions into their wage setting agreements if they do not fall within the network of the majority of the population. Additional analysis in Appendix A also reports separate elasticities by individuals' unionization status, as well as the size of the firm that they work in, in order to understand the role that collective bargaining plays in wage flexibility[11].

Appendix B extends this analysis by estimating local unemployment rates specific to various racial networks, to probe whether cross-racial effects are present: do blacks respond to white labour market conditions and vice versa? Furthermore, the aspect of geographic reach is incorporated into the cross-racial analysis by estimating elasticities using spatially lagged unemployment rates at various distances. While it would be preferable to construct spatially lagged unemployment rates using fine-grained regional definitions, the data does not allow this, and the unit of analysis remains the magisterial district for this section of the analysis.

[10] Race and gender are obvious classifications, as they distinguish very separate groups within the labour market. However, as Baltagi et al. (2012) show for Turkey, skills, age and education may present appropriate categorizations. Younger, less educated workers there are more responsive to high unemployment, likely due to their limited bargaining power. Kingdon & Knight (2006a) also present separate estimates for the black sub-population, as well as various demographic groups. While each of these heterogeneous effects is potentially very real in the South African context, the separation of effects by sub-group is limited to two additional cases in appendices. Appendix A considers differences by union membership and firm size, while appendix B explores responses to unemployment rates of own races and other races.

[11] Bargaining councils are not measured directly as is done by Bhorat et al. (2012). Only union status is contained in the data. Given the larger role for sectoral agreements in wage determination in South Africa than for unions directly, this proxy for collective bargaining is poor.

Despite the fact that indicators are constructed so that spatially distant units are down-weighted, the estimation of unemployment rates by demographic network shows how racial ties diminish the effects of "long distance unemployment" in creating wage setting relations that stretch beyond individuals' immediate locations.

A priori expectations suggest that optimal local labour markets may be defined at the district council level (and in some cases magisterial districts), for the simple reason that bargaining councils are defined by these boundaries, and have been effectively used in the identification strategy of labour market effects elsewhere (Magruder, 2012).

2.5. Results

The discussion commences with results from Table 2.2 that highlight various racial groups' wage setting responses to local unemployment rates of the entire population in their region (where locality is defined at three levels: magisterial districts, district councils and provinces). These results illustrate three features: firstly, that size of the region at which unemployment is measured matters; secondly, that accounting for unobserved time-invariant spatial heterogeneity is influential in the South African context; and thirdly, that not all groups are equally sensitive to overall labour market conditions within their regions. In addition, the role of reverse causality and measurement error is investigated with instrumental variables' estimates in later sections.

2.5.1. Optimal labour market size

Starting with the first observation, column 1 of Table 2.2 estimates the responsiveness of individual wages to overall broad unemployment rates within magisterial districts, with separate elasticities estimated by race and gender[12]. No geographic fixed effects are included. Since magisterial districts are the smallest geographic units observed in labour market data, it is also the smallest possible labour market definition that can be delineated by this study. While Figure 2.1 highlights that broadly speaking, adjacent magisterial districts have similar unemployment rates, in some cases aggregating to larger geographic regions would potentially introduce measurement error of the kind that is analysed in equation (2.9) above. The discussion first focuses on the black population group. Kingdon & Knight (2006a), without appropriate spatial fixed effects, found a wage curve for the population as a whole, with a slightly weaker estimate for the black sub-population[13]. For black males, an estimate of

[12] Analysis in Appendix B differentiates how various groups respond to regional unemployment rates that are delimited to race and gender groups.
[13] In results not shown, the specification of choice does not render a wage curve for the population as a whole, with or without spatial fixed effects. This is also reflected in the lack of wage curve behaviour for the other population groups.

-0.076 is statistically different from zero at the 10 per cent level of significance, and is close to the standard magnitude of -0.1 found in the international literature. The elasticity for black females is only slightly smaller, but statistically insignificant. However, both genders among this group have elasticities that are not statistically different from -0.1, while this proposition is rejected for all other groups.

Moving to column 5, where unemployment is estimated at the district council level (but no fixed effects are accounted for), the wage curve estimates are about twice the magnitude in absolute value for the black population group, though they are also less precisely estimated. While the coefficients are not statistically different from zero at conventional significance levels, they are also not statistically distinguishable from -0.1 for the black population group. In sum, moving to the district council level has raised wage curve estimates, but also raised standard errors. Because wage curve magnitudes are more negative when using district council unemployment rates, predictions from equation (2.9) suggest that this demarcation is a more appropriate labour market definition than magisterial districts for the black population group.

In column 8, where provincial unemployment rates are used in estimation (without any fixed effects), the magnitude of wage curve estimates for the black population group turns positive. Importantly, moving to this larger regional demarcation shifts the coefficient away from a large negative magnitude. This can be explained by measurement error that is introduced by choosing a labour market definition that is too large in geographic reach. When unemployment rates are calculated for geographically large regions (but where sub-regions display large levels of heterogeneity within these borders), coefficient sizes move away from the norm of -0.1 towards zero, as equation (2.9) predicts[14]. Given that a maximally negative value is obtained for estimates using district council unemployment rates, this appears to be the most optimal labour market definition that can be discerned in the data. Both smaller and larger labour market demarcations yield less negative estimates; one reason is that measurement error from sub-optimal labour market demarcations is present. Estimates using magisterial district unemployment rates are, however, more efficient and yield wage curve behaviour for the black population group.

2.5.2. Spatial heterogeneity

Turning to the second concern, the importance of spatial heterogeneity is also clearly illustrated in the South African context. For each of the specifications, unobserved spatial

[14] Equation (2.9) predicts that the coefficient would tend to zero and not necessarily switch to a positive value, so that factors other than labour market definition are also at play here. Nevertheless, these estimates indicate that provinces do not define labour markets well for the black population group.

heterogeneity is accounted for, starting with rough fixed effects for provinces, and progressing to those for smaller regions (district councils and magisterial districts). Column 2 is akin to the estimation of Kingdon & Knight (2006a), where magisterial district unemployment rates are used with provincial fixed effects. Notably, the point estimate hardly changes for the black population group, though it is more efficiently estimated relative to column 1. These results, using similar methods, align very closely with the estimates for 1993 by Kingdon & Knight (2006a), where a wage curve elasticity of -0.072 is found for black South Africans. Given the concerns highlighted in equations (2.6) and (2.7), it is however potentially necessary to account for fixed effects of geographically smaller regions. Column 3 does so using district council fixed effects, raising their number from 9 to 55. Again, coefficient estimates are not substantially influenced. At this juncture one might conclude that black South African wage setters are highly responsive to local unemployment rates; furthermore one might add that the conclusions of meta-analyses (Nijkamp & Poot, 2005), that fixed effects do not alter conclusions, apply in this instance also.

However, the multiple time periods in the data allow for the estimation of column 4, where magisterial district fixed effects can be included alongside unemployment rates estimated at the same level. The coefficient for the black population becomes statistically insignificant, and is now also distinguishable from -0.1 at the 1% level. In other words, black South Africans do not exhibit transient wage curve behaviour once appropriate controls for spatial heterogeneity are introduced; rather, the negative relationship may indicate a long-term spatial equilibrium, rather than short-run responses of wages to slack local labour markets. In light of equation (2.6), the finer fixed effects reduce the inconsistency that rougher fixed effects cannot. In fact, these estimates concur with the assertions that wages are inflexible, and that labour market conditions do not influence how demands are formed (Fedderke, 2012). Rather, one might argue, collective bargaining and unionization lead to simultaneously higher wages and unemployment, and an inflexible labour market (Magruder, 2012). Results in Appendix A explore this proposition further, showing that the null effect that is observed here is dominated by union members in larger firms. Non-unionized workers are responsive to local labour market conditions, though they are less sensitive to local unemployment when they work in large firms. This result indicates that collective bargaining agreements are extended to these types of workers, and contributes to wage inflexibility.

A similar exercise is repeated for unemployment rates estimated for larger regions. Column 6 introduces provincial fixed effects when unemployment is estimated at the district council level, with inconsequential changes in the wage curve magnitude. However, once district council fixed effects are introduced in column 7, the previously significantly negative estimate for blacks collapses to a magnitude that is statistically indistinguishable from zero.

This result is generally true for all demographic groups, except for the white population, which has a robustly large positive elasticity. Notwithstanding, the coefficients for black males is ten times the size of estimates using magisterial district unemployment rates with appropriate fixed effects. Provincial fixed effects - when estimating unemployment rates for the same region type (in column 9) - do not alter magnitudes substantially, but this is from a baseline estimate in column 8 that was already positive. In sum, finer fixed effects tend to reduce the inconsistency in wage curve estimates in South Africa for the black population, and provincial unemployment rates are inadequate at capturing local labour market conditions. A transitory wage curve is unlikely in all specifications, though inefficient estimation could be clouding this result.

Up to this point, only wage curve estimates for the black population group were analysed. This is because this group constitutes a majority of the population, and also dominates the overall unemployment rate used in these estimates. Interestingly, other groups are not as responsive in their wage setting to the overall unemployment rate within their regions. In most cases, estimates are statistically insignificant, except for whites (and in some cases coloured males). Regardless of specification, estimates for the white population group are positive and statistically significant in most instances. They are, however, always statistically distinguishable from -0.1. Does this constitute a deviation from wage curve behaviour? At first glance it does, and one might falsely conclude that white individuals raise their wages in defiance to slack labour market conditions, asserting a strong bargaining position. However, these results rather suggest that white individuals constitute a distinct labour market, and that they potentially temper their wages only when unemployment is high within their own ethno-demographic network. This eventuality is investigated in greater detail in Appendix B. Without any spatial fixed effects, whites wages are significantly negatively correlated with their *own* group's unemployment rate, suggesting that whites do constitute a separate labour market. However, once introducing appropriate fixed effects, elasticities are insignificant, so that white individuals are generally non-responsive to local labour market conditions of all groups.

2.5.3. Measurement error and reverse causality

The fixed effect analysis does not account for measurement error explicitly, nor for reverse causality. Table 2.3 presents instrumental variables' estimates, with and without spatial fixed effects[15]. Results without fixed effects are discussed first, in order to understand only the role

[15] Instead of a two-stage least squares (2SLS) estimator, a control function approach is followed. In the first stage unemployment model, no differentiation by race and gender is introduced. The second stage contains a residual from the first, but allows the instrumented variable to be interacted with

of measurement error and reverse causality. Specification 10 follows the standard in the literature, by using the time lagged log unemployment rate as an instrument for contemporaneous local unemployment. Instruments are strong, with a first stage F statistic of 183.81, and the Hausmann test concluding that results differ significantly from the OLS estimates. As is standard, instrumentation increases the absolute value of the wage curve elasticity, indicating higher wage flexibility (Baltagi et al., 2012; Kingdon & Knight, 2006a). While it is possible that reverse causality is eliminated, it is also likely that measurement error associated with incorrectly chosen demarcations is accounted for. For instance: the estimates for black males and females in specification 10 in Table 2.3 (using instrumented magisterial district unemployment rates without fixed effects) are very close to comparable estimates in specification 5 in Table 2.2 (using district council unemployment rates without instruments or fixed effects). Hence, it is possible that the instrumentation strategy corrects for using a labour market demarcation that is too small (magisterial district), yielding estimates that would have arisen had a more natural labour market definition been chosen (district council).

Using the homeland border discontinuity as an instrument yields even larger estimates, with wage curve behaviour now emerging for all race and gender groups (specification 11). The spatially lagged unemployment instrument provides a comparable specification in column 12. Similar identification strategies, using district councils as unemployment regions, also yield large wage curve elasticities in columns 15 and 16. Geographic instruments are therefore a potentially fruitful alternative to time lags.

Results from instrumental variables' estimates show greater wage flexibility than OLS do. This clearly illustrates that measurement error and reverse causality are influential in South Africa, so that results may be understated. However, the instruments have removed not all forms of endogeneity, as section 2.5.2. illustrates the importance of spatial heterogeneity for estimates. For instance, contemporaneous and lagged unemployment rates both contain a common time persistent, structural component. Should this be correlated in any way with wage fixed effects, IV estimates remain inconsistent. It is likely that long-run wages are related to long-run unemployment, so that corrections for spatial heterogeneity are required in IV estimates of this kind also.

Specifications 13, 14, 18 and 19 instrument *and* introduce appropriate fixed effects for magisterial districts and district councils respectively. Time and spatial lags are used as instruments, though the homeland border discontinuity is discarded due to its time-invariance and perfect collinearity with spatial fixed effects. For both demarcations, time lags yield large

---

demographic classifications. While standard errors should be adjusted to be more conservative, this is not implemented here (Imbens & Wooldridge, 2007).

negative (but statistically insignificant) wage curve estimates. The instruments are, however, particularly weak, with first stage F statistics below 1. As a result, these large, inefficient estimates are biased and inconsistent. Given that the first and second stages are stripped of the time persistent component of unemployment by the introduction of fixed effects, the correlation between the transient components in consecutive periods is poor. While other studies successfully use this instrumentation strategy with fixed effects (Baltagi et al., 2012), it is not useful when local unemployment is structural and permanent in nature, with little of the variation occurring across *time*.

Alternatively, specification 14 instruments with the *spatial* lag. Variation across space remains intact after fixed effects remove individual means over time. Instruments are now strong, with a first stage F statistic of 22.99. The Hausmann test reveals that introducing instruments to the fixed effects specification has, however, *not* changed estimates compared to specification 4, so that wages remain inflexible. Elasticities are either statistically insignificant or positively significant for the various demographic groups. The same estimate using district council unemployment rates again yields large and insignificant elasticities, though the instrument is again weak.

In all cases when instruments are weak, elasticities are overstated; whenever they are strong, elasticities are closer to zero once fixed effects are introduced. While the potential exists that the spatial lag is nevertheless endogenous to wages[16], specification 14 is the best estimate that attempts to solve for each of spatial heterogeneity, reverse causality and measurement error. Even without instrumentation, spatial fixed effects reduce the wage curve elasticity to zero. The implication is that wages are not responsive to local unemployment in South Africa, concurring with the sentiments of Fedderke (2012), but rejecting earlier evidence by Kingdon & Knight (2006a) on econometric grounds.

[16] If wage setters actually take larger regions' unemployment rates into account when bargaining, the instrument will be correlated with wages *directly*, rather than indirectly through own-district unemployment. As a result, the exclusion restriction may be violated. Appendix B illustrates this possibility.

Table 2.2 OLS wage curve estimates with various regional unemployment rates and fixed effects specifications, by race and gender

Dependent variable: *log(monthly earnings)* - OLS		1	2	3	4	5	6	7	8	9
	Unemployment region	Magisterial district (x354)				District council (x55)			Province (x9)	
Wage curve elasticity: Coefficient on *log(broad unemployment rate)* by group	Black male	-0.076*	-0.082**	-0.074**	-0.004	-0.148	-0.114*	-0.041	0.089	0.173
		(0.040)	(0.033)	(0.032)	(0.029)	(0.095)	(0.063)	(0.069)	(0.166)	(0.152)
	Black female	-0.067	-0.075**	-0.067**	-0.028	-0.104	-0.086	-0.027	0.014	0.106
		(0.041)	(0.038)	(0.034)	(0.038)	(0.098)	(0.092)	(0.097)	(0.145)	(0.137)
	Coloured male	0.029	0.081*	0.071**	0.062*	-0.141**	0.013	0.071	0.174*	0.348*
		(0.043)	(0.042)	(0.035)	(0.033)	(0.068)	(0.072)	(0.076)	(0.083)	(0.181)
	Coloured female	0.014	0.061	0.051	0.043	-0.188**	-0.037	0.021	0.067	0.243
		(0.059)	(0.060)	(0.051)	(0.048)	(0.093)	(0.090)	(0.090)	(0.113)	(0.206)
	Indian male	0.101	0.075	0.071	-0.025	0.172	0.140	0.207	-0.103	0.010
		(0.075)	(0.062)	(0.062)	(0.073)	(0.138)	(0.135)	(0.149)	(0.166)	(0.176)
	Indian female	0.103	0.076	0.069	-0.038	0.115	0.069	0.137	-0.172	-0.067
		(0.073)	(0.059)	(0.053)	(0.053)	(0.236)	(0.213)	(0.215)	(0.285)	(0.312)
	White male	0.106**	0.128***	0.128***	0.146***	0.078	0.159**	0.218***	0.264	0.423**
		(0.048)	(0.041)	(0.039)	(0.039)	(0.078)	(0.070)	(0.078)	(0.146)	(0.153)
	White female	0.128***	0.160***	0.155***	0.169***	0.114	0.212***	0.271***	0.326**	0.489**
		(0.039)	(0.035)	(0.036)	(0.038)	(0.070)	(0.065)	(0.076)	(0.130)	(0.193)
	Age	0.107***	0.107***	0.106***	0.104***	0.108***	0.108***	0.107***	0.110***	0.110***
		(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)	(0.007)	(0.007)
	Age^2	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	Education	-0.030***	-0.029***	-0.030***	-0.030***	-0.025***	-0.025***	-0.026***	-0.019*	-0.020*
		(0.004)	(0.004)	(0.004)	(0.004)	(0.006)	(0.005)	(0.005)	(0.009)	(0.009)
	$Education^2$	0.011***	0.010***	0.011***	0.011***	0.011***	0.010***	0.011***	0.010***	0.010***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Period FE & Race x Gender FE & Region Controls		Y	Y	Y	Y	Y	Y	Y	Y	Y
	Province FE (x9)	N	Y	N	N	N	Y	N	N	Y
	DC FE (x55)	N	N	Y	N	N	N	Y	N	N
	MD FE (x354)	N	N	N	Y	N	N	N	N	N
	Constant	3.581***	3.424***	3.447***	3.656***	3.059***	2.488***	2.800***	4.939**	1.527**
	R-squared	0.538	0.543	0.547	0.556	0.535	0.539	0.542	0.529	0.53
	N	186353	186353	186353	186353	186731	186731	186731	186731	186731

NOTES: * p<0.1, ** p<0.05, *** p<0.01. Own calculations from LFS 2000b to 2004a. Monthly earnings measured at individual level and deflated by national CPI. Unempl measured by broad definition, at geographic level indicated in headings. Additional controls for *regional education composition, regional occupation composition, regional sector composition*. Standard errors, clustered at geographic level at which unemployment is measured, in parentheses.

Table 2.3 Instrumental Variables wage curve estimates with various regional unemployment rates and fixed effects specifications, by race and gender

Dependent var: *log(monthly earn)*		10	11	12	13	14	15	16	17	18	19
Unemployment region		Magisterial district (x354)					District council (x55)				
Instrumental variable		Time lag	Border	Spatial lag	Time lag	Spatial lag	Time lag	Border	Spatial lag	Time lag	Spatial lag
Wage curve elasticity Coefficient on *log(broad unemployment rate)*	Black male	-0.141**	-0.329***	-0.320**	-0.828	0.027	-0.294**	-0.496***	-0.046	-0.230	-0.809
		(0.063)	(0.093)	(0.141)	(3.344)	(0.089)	(0.120)	(0.163)	(0.349)	(1.142)	(2.541)
	Black female	-0.144**	-0.324***	-0.316**	-0.859	0.003	-0.263**	-0.445***	-0.002	-0.226	-0.795
		(0.057)	(0.088)	(0.134)	(3.347)	(0.090)	(0.108)	(0.165)	(0.315)	(1.145)	(2.529)
	Coloured male	-0.036	-0.236***	-0.258*	-0.759	0.073	-0.328***	-0.384***	-0.038	-0.141	-0.697
		(0.063)	(0.083)	(0.132)	(3.344)	(0.081)	(0.096)	(0.134)	(0.335)	(1.168)	(2.539)
	Coloured female	-0.060	-0.274***	-0.294**	-0.784	0.035	-0.371***	-0.422***	-0.084	-0.192	-0.746
		(0.078)	(0.085)	(0.136)	(3.346)	(0.081)	(0.121)	(0.144)	(0.350)	(1.158)	(2.553)
	Indian male	0.024	-0.142*	-0.131	-0.859	0.009	0.056	-0.124	0.274	0.043	-0.560
		(0.091)	(0.083)	(0.155)	(3.352)	(0.104)	(0.190)	(0.136)	(0.332)	(1.182)	(2.535)
	Indian female	0.022	-0.143*	-0.131	-0.880	-0.005	-0.082	-0.179	0.219	-0.107	-0.630
		(0.090)	(0.081)	(0.150)	(3.355)	(0.093)	(0.240)	(0.229)	(0.356)	(1.197)	(2.546)
	White male	0.067	-0.107	-0.135	-0.657	0.179**	-0.020	-0.163	0.177	0.065	-0.550
		(0.071)	(0.103)	(0.137)	(3.354)	(0.090)	(0.094)	(0.151)	(0.335)	(1.151)	(2.534)
	White female	0.052	-0.077	-0.107	-0.668	0.212**	-0.048	-0.100	0.213	0.061	-0.497
		(0.051)	(0.095)	(0.135)	(3.345)	(0.092)	(0.088)	(0.146)	(0.333)	(1.160)	(2.556)
Period FE		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Race x Gender FE		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Other Controls		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
DC FE	(x55)	N	N	N	N	N	N	N	N	Y	Y
MD FE	(x354)	N	N	N	Y	Y	N	N	N	N	N
Constant		3.471***	2.576***	2.588***	2.394	3.693***	2.243**	1.566	3.598*	2.466	0.528
R-squared		0.542	0.547	0.546	0.561	0.563	0.541	0.544	0.535	0.547	0.542
N		161843	173362	172618	161843	172618	162389	173740	186731	162389	186731
p-value of Hausmann test		0.063*	0.001***	0.049**	0.805	0.731	0.057*	0.048**	0.749	0.866	0.768
F statistic of first stage		183.81	30.908	47.31	0.044	22.99	47.532	27.713	14.139	0.031	0.005

NOTES: * p<0.1, ** p<0.05, *** p<0.01. Own calculations from LFS 2000b to 2004a. Monthly individual earnings deflated by national CPI. Broad Unempl measured at geographic level indicated in headings. Additional controls: *Age*, *Age*^{2}, *Educ*, *Educ*^{2}, *region educ composition*, *region occup composition*, *region sector composition*. Standard errors, clustered at geographic level, in parentheses. Control functions do not adjust standard errors in 2nd stage. Border IV measures straight line between regional centroid and closest former apartheid homeland border, with a discontinuity at the border. Spatial lag IV represents the spatially weighted unemployment rate within a radius of 200km of regional centroid. Time lag IV represents previous period's unemployment rate within same region.

2.5.4. Long-run equilibrium

South Africa's unemployment problem is structural, so that long-run factors are salient in its analysis (Bhorat & Hodge, 1999). Accounting for fixed effects separates the long-run equilibrium from the transient wage curve. The high wage flexibility recorded by Kingdon & Knight (2006a) may confound a true zero transient effect with a long-run trade-off between wages and unemployment, in much the same manner as in Nordic countries (Albaek et al., 2000). This section investigates this possibility.

Figure 2.4 plots the wage fixed effects (that are extracted from various specifications) against magisterial district unemployment rates[17]. The fixed effects represent the long-run component of wages. Coefficients that are estimated (with the inclusion of fixed effects) in Tables 2.2 and 2.3, however, represent short-run relationships. Differences between coefficient estimates and the relationships depicted in Figure 2.4 therefore emphasise the importance of separating long-run relationships from short-run wage curve estimates. Specification 4 in Table 2.2 (with no instruments) yields fixed effects that are negatively related to unemployment. Similarly, specification 12 in Table 2.3 (with spatially lagged instruments) yields a negative long-run relationship. In both of these cases, a null effect was found for transient wage curve estimates. In contrast, specification 10 in Table 2.3 yielded an insignificantly negative (but biased) short-run elasticity, with a positive relationship in the long-run shown in the figure. The most trustworthy estimates therefore point to zero short-run flexibility, but a long-run trade-off between regional unemployment and wages. Appendix C shows that the long-run elasticity associated with the non-instrumented results in Figure 2.4 is remarkably close to -0.1[18].

Biased estimates render a short-run wage curve with a long-run spatial equilibrium that would agree with migration models in the Harris & Todaro (1970) tradition. Kingdon & Knight (2006a) therefore find a solution that confounds the short-run inflexibility of wages with a long-run equilibrium that reflects migration to regions with relatively high wages and low unemployment.

[17] This demarcation is used because all instruments were weak when using district councils.
[18] Appendix C also attempts to isolate long-run historical and geographic factors that contribute to this long-run trade-off. While these influences wages, only current firm structure within districts alters the long-run relationship to any degree.

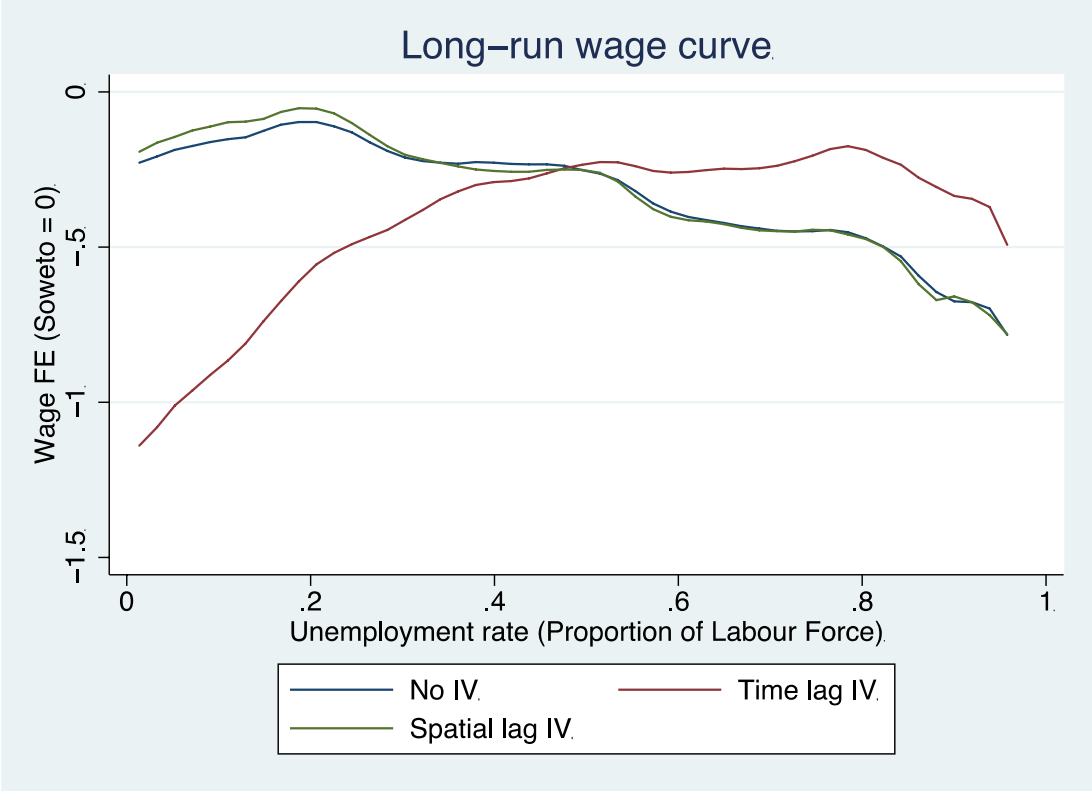

Figure 2.4 Relationship between wage fixed effects and broad local unemployment rates

NOTES: Own calculations from Labour Force Survey 2000b to 2004a. "No IV" uses fixed effects from specification 4 in Table 2.2; "Time lag IV" uses fixed effects from specification 10 in Table 2.3; "Spatial lag IV" uses fixed effects from specification 12 in Table 2.3. Unemployment is measured by the broad definition, for the whole population, at the magisterial district level

2.6. Conclusion

While many studies agree that South Africa has a rigid labour market together with inflexibility in wage determination, one piece of existing microeconometric evidence has concluded the opposite (Kingdon & Knight, 2006a). However, this chapter has highlighted the importance of correctly defining local labour markets, accounting for spatial heterogeneity and using instrumental variables that are not reliant on time variation. The standard approach of using time lagged instrumental variables is not appropriate in a country where local unemployment is persistent. This study proposes the use of spatially lagged variables to identify the wage curve. Each of these microeconometric solutions removes bias of a different variant. In sum, adjusting for shortcomings yields a metric that points to an inflexible labour market, providing an updated view of previous evidence.

Firstly, this paper has shown that in the short-run wages are not very responsive to local labour market conditions, once long-run effects are netted out by appropriate spatial fixed effects. While instrumentation (in order to account for using inappropriate labour market demarcations and reverse causality) raises the magnitude of elasticities, combining this with

fixed effects again yields a zero relationship. This concurs with macroeconomic evidence, suggesting that South African wage setting does not appear to respond to slack labour market conditions over time.

The best local labour market definition in survey data appears to be the district council. Two potential reasons can be offered for this finding. Firstly, smaller magisterial districts are not adequately sampled in household surveys and unemployment rates therefore lead to attenuated wage curve estimates. However, instrumentation successfully alleviates this problem. Secondly (and more plausibly), district councils are more appropriate, because collective bargaining agreements often cover regions that are larger than magisterial districts, but smaller than entire provinces.

Additional results suggest that the overall null effect is dominated by unionized workers, particularly those that work in larger firms. This matches the notion that collective bargaining agreements extend negotiations to entire regions and industries, limiting the adjustment of wages when labour market conditions are slack. In particular, respondents working in small firms exhibit short-run responses to local unemployment. Non-unionized workers as a whole are more flexible in wage setting.

This is not to say that there is no trade-off between wages and local unemployment. Long-run local wages are correlated with local unemployment, in compliance with usual wage curve behaviour when the preferred specification is investigated. Hence, the long-run spatial equilibrium consists of high wage-low unemployment (urban areas), though some former homeland regions still pay (relatively) high wages in the face of high unemployment. Estimates that do not separate long-run from short-run scenarios confound two time horizons, reflecting only the long-run equilibrium, but falsely concluding that there is flexibility in the short-run.

South Africa does not follow the typical developing country pattern due to its high unemployment rate (Fields, 2011), and because its long-run spatial equilibrium is not one of high wage-high unemployment urban regions (Harris & Todaro, 1970). Nor does it strictly comply with the international norm of short-run wage flexibility in response to slack labour market conditions (Blanchflower & Oswald, 2008) due to centralized bargaining institutions (Albaek et al., 2000). The short-run rigidities found by this paper suggest that the labour market is unlikely to adjust to high unemployment through wage adjustment. As a result, migration is unlikely to occur to regions with high unemployment, which moderates long-run wage demands in those areas. This provides one potential explanation for the relatively low levels of migration following the removal of spatial restrictions by the post-apartheid regime.

Chapter 3

Errors in recalling childhood socioeconomic standing: the role of anchoring and household formation

3.1. Introduction

The central influence of early life circumstances on outcomes later in life is a link that economists now recognize widely (Almond & Currie, 2011). Diverse life choices are enabled or limited by the environment in which children grow up, ranging from early cognitive development to schooling choices and reaching as far as labour market prospects late in the life cycle.

Longitudinal life course surveys provide the ideal source of data to investigate the relationship between childhood inputs and later life outcomes. However, this data has to be collected over long periods of time; the surveys are typically very expensive to administer; and this data therefore is often not available in developing countries. In the absence of longitudinal data, retrospective reports of earlier life circumstances provide an alternative means of assessing the influence of childhood on adult outcomes. For example, retrospective data is used when studying the effect of childhood socioeconomic position on adult health outcomes (McKenzie & Carter, 2009).

In this study, we evaluate the reliability of retrospective reports of late childhood (adolescence) socioeconomic status (SES), using data collected in two waves of a longitudinal household survey. The survey was conducted in South Africa, a developing country with high inequality and low levels of social mobility, both within and across generations (Piraino, 2014; Lechtenfeld & Zoch, 2014; Finn et al., 2014), and therefore a country where childhood circumstances should have a long reach on diverse outcomes such as schooling and labour market prospects. We consider whether various reports of late childhood socioeconomic status, collected from the same adults in consecutive waves of the survey, are affected by *current* demographic and economic characteristics of the respondent, as opposed to representing true reflections of the past. We evaluate two measures of childhood socioeconomic status for this type of anchoring effect: the education of the adult's parents and the adult's recall of the household's economic status at age 15.

Childhood socioeconomic status among adults is most commonly measured using information on parental (and typically paternal) occupation or education at the time of the adult's childhood. Importantly, this information is reported by respondents in their adult lives, and not by the parents themselves. A number of studies that assess these particular retrospective reports generally find

evidence supporting their reliability (Berney & Blane, 1997; Kriegler et al., 1998; Batty et al., 2005; Ward, 2011). However, in developing countries such information may be less useful. In South Africa, for example, large numbers of children grow up in single-parent households, typically headed by mothers. This is a consequence of low marriage rates and a migrant labour system, implemented under the apartheid regime, where many fathers continue to work and live in urban regions away from their families (cf. Hill et al., 2008; Posel & Rudwick 2013). As a result, information on father's education or occupation is often not reported or may be irrelevant. Similarly, adults may have difficulty recalling mother's education[19]. Arguably, these indicators are also more useful to establish the extent of *intergenerational* rather than *life-course* mobility. This is because parental education directly represents a previous generation's endowments, while it only indirectly influences socioeconomic status during childhood.

An alternative measure of childhood socioeconomic position is that provided by retrospective (subjective) appraisals of economic status. For example, respondents may be asked to rate the economic status of their household at some point in childhood, from 'very poor' to 'rich', or relative to the economic status of other households. Only a few studies verify the usefulness of such data. Ward (2011) and Straughen et al. (2013) compare consistency in subjective appraisals of childhood socioeconomic status across different, but closely related individuals (siblings, or mothers and daughters). The studies find low concordance between intra-family reports of socioeconomic status, but this is at least partly because subjective assessments are provided by different people.

In this study, we use replicate retrospective reports on childhood economic status, collected from the same adults in consecutive waves of a South African panel study, to assess the reliability of recall data on economic status during childhood. In particular, we investigate whether there is evidence of anchoring in retrospective reports, whereby changes in current circumstances are systematically associated with changes in retrospective reports of the same individual over time (Haas, 2007). Tversky & Kahneman (1974) define anchoring as a cognitive process, whereby individuals assess their current circumstances and then obtain their view of another period by applying an incremental change to this initial anchoring point. Hence, the past can be interpreted incorrectly in light of assessments of current circumstances. The longitudinal approach followed here also allows us to control for time-invariant heterogeneity (such as an innate ability to remember past events and other personality characteristics) using fixed effects estimation. Intra-family studies that are typically used

[19] Moreover, it is not clear what domain of childhood is measured by parental attributes. Harper et al. (2002), for instance, suggest that parental education measures the child's intellectual environment, while parental occupation measures the material environment during childhood. Each is a measure of one facet of childhood socioeconomic status. The former is potentially more suitable to assess the path of education decisions over the life cycle, while the latter attribute affects the financial position of households and credit constraints that influence future choices.

to gauge reliability of retrospective data are not able to account for these person-specific omitted factors.

Finally, the paper discusses the implications of anchoring for empirical analysis. Measurement errors usually lead to biased estimation of regression coefficients. As a result, anchoring compromises causal estimation of childhood reach. While other authors have used repeated measures of the same assessment to conduct instrumental variables' estimates (Crossley & Kennedy, 2002), this is also not a solution when anchoring effects are persistent. Additionally, this has implications for the literature on household bargaining, which uses recalled childhood background characteristics for identification (Browning & Bonke, 2009; Browning & Lechene, 2003).

In the next section, we review the literature on childhood recall and in section 3 we describe the data analysed in the study and our methodology. Section 4 presents the results and a discussion of the implications of poor recall for microeconometric analysis, while section 5 concludes and summarizes the main findings.

3.2. Review: the use of retrospective and subjective data

Retrospective data is potentially useful to explore the long reach of childhood in adult life. Most of the research on the dependence of life cycle outcomes on childhood circumstances focuses on developed country populations (Almond & Currie, 2011; Cunha et al., 2006). This is at least partly because of the availability of long-established surveys that track individuals from young ages into the labour market, for instance the National Longitudinal Survey of Youth (NLSY) in the United States. Conducting such long-run studies incurs high monetary costs because individuals need to be tracked or followed, which involves potentially large distances if participants are spatially mobile. Attrition is also likely to be severe in such a scenario. Furthermore, the commencement of such a study with a cohort of youths will only yield long-run outcomes at a much later date, implying a long "gestation period" for conclusions to be drawn. These difficulties entail that birth cohort studies are scarce in developing countries. Yet, given the potential for chronic poverty and low levels of income mobility, it is precisely in these regions that an understanding of the reach of childhood factors is particularly important.

In the absence of life course longitudinal data, recall data is a potentially important alternative. In contrast to birth cohort studies, retrospective data is "available quickly" (Haas 2007:115), and questions asking adults to recall socioeconomic conditions during childhood are easily included, and at relatively low cost, in existing surveys. However, the usefulness of this data requires that adults are able to provide a reliable recall of their past, and that this recall is not substantially influenced by the current characteristics or circumstances of the individual.

In their study of the quality of retrospective data in a developing country setting, Becket et al. (2001) describe four typical patterns of errors in recalling the past. The first concerns the length of recall, whereby the longer the period of recall, the more likely the past will be remembered with error. We would therefore expect childhood to be recalled with greater error among older adults. However, where replicate retrospective reports are provided by the same adult, we would not expect inconsistency in these reports to derive from length of recall errors, particularly if the retrospective assessments are collected in relatively close time proximity to each other. In this case, length of recall errors should be strongly correlated over time, with a consistent error surfacing at both reporting occasions.

The second type of recall error concerns the salience of what is being recalled. Circumstances and events will be recalled with less error if they are more prominent in an individual's life: life-changing events, or particularly severe or opportune circumstances, are likely to be remembered with less error. Hence, childhood socioeconomic status may only be particularly memorable (and recalled correctly) if adults lived in extreme poverty or wealth as children. However, severe childhood circumstances may also result in non-random non-response among adults, particularly when these circumstances are associated with instability and frequent changes in the family's living arrangements (McKenzie & Carter, 2009).

Third, people's recall of the past may be influenced by telescoping, whereby they estimate the timing of past events incorrectly. In particular, memorable events may be recalled as having occurred more recently than they did. In this case, adults who experienced distinct upward (or downward) mobility during childhood may remember this transition as having occurred more recently (for example, during the early stages of adulthood). Telescoping therefore may result in life-time mobility between childhood and late adulthood being overstated.

The fourth type of error in recall undergirds each of the others, and concerns the "accessibility of events" (Brown et al., 1986): occurrences that are easier to recall are reported as having occurred more frequently. In the case of childhood socioeconomic status, if many temporary shocks to consumption would have affected a household, an individual might recall more permanent depressed economic circumstances, despite their non-permanence.

A further problem that affects the reliability of recall data is related to anchoring, a cognitive bias which was not investigated by Becket et al. (2001). Anchoring occurs when retrospective reports are influenced by the current circumstances or status of the respondent (Haas, 2007). For example, adults' assessments of socioeconomic status during childhood will be anchored by their current socioeconomic status if they project these circumstances backwards. Tversky & Kahneman (1974) emphasize that under uncertainty individuals tend to make judgments based on beliefs that are simplified by a number of heuristics. While simplification reduces the cost of making fairly good

judgments, errors do arise. Anchoring entails that individuals choose a starting information set, from which adjustments are made to conclude on another reference point. In the case of assessing past socioeconomic status, individuals first consider their current position, and then adjust backwards by an increment which reflects their perceived rise or fall in the income distribution between the starting and end point. Most experiments show that the starting point is influential for the final estimate, even if the degree of adjustment could be correct (Tversky & Kahneman, 1974). Hence, if current circumstances are out of the ordinary, estimates of the past will be biased in the same direction. If anchoring occurs, however, then the usefulness of retrospective data in exploring the reach of childhood is greatly compromised.

The literature that evaluates the reliability and validity of retrospective data is relatively limited, and there are even fewer studies from developing countries (where such recall measures would be particularly valuable in the absence of birth cohort studies). Moreover, the collection of replicate retrospective reports in longitudinal datasets is not common. Survey designers often choose to "benchmark" to reduce interview time (Beckett et al. 2001: 622). That is, interviewers first remind respondents about information that they provided in the previous round, so that a relevant reference point is established for subsequent answers. Consequently, different reports from the same individual are not collected, and few studies can therefore test for evidence of anchoring in retrospective data.

One exception is a study by Haas (2007), which evaluates reports on retrospective child health status in a developed country. The study finds that ordinal retrospective reports (across two waves of the Panel Study of Income Dynamics in the United States) are exactly consistent for approximately 55 per cent of respondents, with some marginal shifts to adjacent categories also prevalent in the data. In order to eliminate these fine differences at category thresholds, measures are also dichotomized, elevating consistency to beyond 90 per cent[20]. Strong consistency suggests that childhood health is a salient characteristic that is recalled accurately by adults. The study also finds that evidence for systematic differences in length of recall is weak, with older individuals recalling their childhood as consistently as younger respondents. However, educated and white individuals were more consistent in their responses over time. Investments in human capital therefore brought about a potential learning effect.

A few studies assess the validity of retrospective data on socioeconomic status specifically. Some research compares adults' recall of childhood circumstances with information that had been gathered years earlier, during childhood (Batty et al. 2005; Brown 2012). Other studies evaluate the reliability of retrospective reports by comparing assessments of childhood socioeconomic status provided by different family members – from pairs of adult female twins (Kriegler et al. 1998); siblings (Ward 2011); or mothers and their adult daughters (Straughen et al. 2013).

[20] This is also done to eliminate insubstantial anchoring effects at the margin.

The findings from these studies are not consistent. Batty et al. (2005), for example, report only "moderate" agreement between retrospective reports of fathers' occupation compared with data collected early in life (54 per cent of responses corresponded). Brown (2012:21), in contrast, finds "typically small" differences between adults' assessments of economic status at childhood and information that had been collected during childhood from the children's mothers. However, the likelihood of inconsistent reports was far higher among those living in larger households, the less educated and among those with less stable family backgrounds.

Studies which compare retrospective reports provided by different family members typically find strong concordance when these reports are on father's education or occupation (Krieger et al. 1998; Ward 2011), but far weaker agreement on subjective reports of childhood socioeconomic position (Ward 2011, Straughen et al. 2013). Krieger et al. (1998), for example, compare responses from pairs of adult female twins on parental occupation and education during childhood, collected in a cross sectional survey. Their results show strong concordance in the recall of father's education (91 per cent agreement), and slightly lower concordance in the recall of father's occupational status during childhood (80 per cent). However, there is no evidence that the extent of agreement varied according to the education or social class of respondents.

Ward (2011) also finds strong correspondence between siblings' retrospective reports of childhood socioeconomic status when this is measured using father's education and occupation. However, concordance is far lower for a subjective question on how financial status at childhood compared to that of other families, even when the seven ordinal responses are grouped into broader categories. Similarly, Straughen et al. (2013) identify low agreement between mothers and their daughters in how they assessed the economic status of the family during the daughter's childhood (with five response options, ranging from "very poor" to "well to do") (Straughen et al 2013: 296).

These findings may suggest that subjective measures (or individual perceptions of economic status) are recalled with greater error than more objective measures, and indeed both Ward (2011) and Straughen et al. (2013) advocate the use of objective indicators of childhood socioeconomic status (such as parental education and occupation). However, in both studies, subjective assessments from two different family members are compared, and therefore neither analysis can control for the possibility of person-specific response patterns in these assessments. In contrast to recall of parental education or occupation, however, various family members may have different perceptions of childhood economic status, influenced by individual experiences since childhood.

In this study, we take advantage of replicate reports, provided by the same individual in two waves of a longitudinal dataset, to assess the reliability of recall data on perceived childhood economic status. Any changes in reports can be attributed to occurrences between the two surveys, and are not likely to be influenced by events before that. We consider whether differential reports vary systematically

according to the demographic and economic characteristics of the respondent, and we focus on whether there is evidence of anchoring in these reports. In particular, we assess whether changes in current household income and changes in individuals' assessments of their current position in their village income distribution, alter individuals' reports about the past. The object is to understand whether recall answers are coloured by current experiences.

3.3. Data and methods

The data for the study comes from a national panel survey for South Africa - the National Income Dynamics Study (NIDS). We consider data from waves 1 and 2 (collected in 2008 and 2010 respectively), where a series of questions was asked about self-assessed current and retrospective relative economic status of every adult (15 years and older) resident in the household. Additionally, adults were asked to recall the educational attainment of both biological parents. In the first wave 16802 resident adults were surveyed, while this rose to 21613 in the second wave, as new households were incorporated into the survey. 15113 of these respondents were interviewed in both rounds. The spacing of the waves is only two years apart, so that inconsistencies in retrospective reports are not likely to be the result of systematic differences in recall periods since childhood. For similar reasons, telescoping is unlikely to be influential. However, the time between the surveys is long enough so that short-run changes in household or individual circumstances could have led to anchoring among respondents. The models presented below explicitly test for correlates of anchoring.

A first question asked individuals to rate their current household income relative to the distribution of their *village or suburb* in 5 bins, from much below to much above average. While this question could have framed answers to subsequent perceptions, the reference group for all other questions was altered to quiz individuals on *national* rankings, and the scale was also changed to a more abstract categorization. These questions asked individuals to place their household's position on a proverbial ladder with six rungs, the first representing the poorest South Africans and the sixth the richest. Firstly, households were asked to assess this ranking for their household when they were aged 15, a subjective measure of relative adolescent socioeconomic status nationally. A follow-up question assessed the same for respondents' current households. The latter is likely to have been primed by the same ranking at age 15 due to the ordering in the questionnaire. To minimise such framing issues, we use the measure at the local level to capture current subjective income status and the national ranking to capture subjective socioeconomic status at adolescence.

While the assessment of adolescent socioeconomic status may be useful to explore the reach of childhood into later life, one of its weaknesses is that it does not capture circumstances in different periods of childhood (McKenzie & Carter, 2009:23). Furthermore, the subjective nature of the question makes it prone to a number of reporting biases. However, it is potentially easier for

respondents to provide assessments of their own experience than to answer questions about their parents' education and occupation. This factor is particularly relevant in South Africa, where (especially black) children have distinctly low probabilities of living with biological parents (Posel & Rudwick, 2013). In many instances *fathers* are absent because of the migrant labour system that leads many men to work in locations distant from their families; additionally, both marriage and cohabitation rates are low among blacks, so that children are unlikely to live with both parents. Whilst remittances from fathers are nevertheless an important source of income for families, children have limited direct contact with fathers (Madhavan et al., 2014). However, many children also live without mothers, but with grandparents (Duflo, 2003). Household formation patterns raise the potential that individuals do not know as much about their parents' attributes as they would about their own experiences. As a result, response rates on questions about parental education are poor, and one could expect greater levels of uncertainty in responses, leading to lower concordance in responses over time.

The retrospective data is assessed for consistency across the two waves of the panel, by establishing the proportion of identical responses over time along various demographic dimensions. In addition, kappa statistics are presented to measure concordance (Cohen, 1960). These measures adjust for the probability that agreement was achieved purely by chance:

$$\kappa = \frac{P(total\ agreement) - P(agreement\ by\ chance)}{1 - P(agreement\ by\ chance)} \quad \ldots(3.1.)$$

Kappa varies from 0 (perfect discordance) to 1 (perfect concordance). Negative values indicate that observed agreement is less than expected by chance. They rarely occur. However, some estimates below have negative lowerbounds on confidence intervals. Total agreement is simply the proportion of respondents that gave the same answer twice. Chance agreement is defined as the product of respective marginal probabilities at each rating for categories of the variable. However, accounting for the eventuality of random agreement makes these statistics relatively conservative in their conclusions about concordance. Nevertheless, most researchers in social and medical sciences refer to kappa statistics to understand the consistency across ratings. Kappa measures rely on the assumption that assessments are provided independently. In the current application this is not the case, as the same respondent provided the two assessments being compared. This was also the case in the work of Crossley & Kennedy (2002), where concordance of answers for the same respondents was established. Kappa statistics are nevertheless presented for the purposes of comparison with other literature.

Concordance for parental education is measured in categories representing no schooling, primary, incomplete secondary, matric and post-secondary education, which represent important thresholds in individuals' schooling careers. However, where these indicators are used as dependent variables in

models, they are measured continuously. This is because the various groups do not contain the same number of years in each category.

To avoid measuring discordance of the subjective ranking due to small changes across category margins (such as from ladder rung 5 to 6), a broader categorization of childhood socioeconomic status is also investigated, where rungs are considered in groups of two. The grouped measure only has 3 rungs (representing "low" "middle" and "high") as opposed to 6 in the original question. A similar approach is followed with current reports on individuals' position in the village income distribution, where 5 original categories (from "much below average" to "much above average") are at times grouped simply into 3 groups representing "below average", "average" and "above average". While most other studies compare individuals' responses to those of close relatives or administrative data, the analysis presented here studies consistency over time for the *same* individuals to control for person-specific response patterns.

It is not possible to test for the full validity of the data, because we do not have "true data" against which the retrospective measures can be compared (Beckett et al. 2001). However, it is possible to establish whether particular individuals are prone to change their judgments of the same reference point. The key question is whether systematic differences in the reliability of retrospective data surface according to the sociodemographic and *current* economic characteristics of respondents.

Following the descriptive analysis, these propositions are tested using various regressions. Firstly, cross section regressions model the difference in childhood SES assessments across periods as a function of wave 1 characteristics and other changes:

$$\Delta SES_{child;irt} = \beta_0 + \beta_1 \Delta SES_{current\ local;irt} + \beta_2 SES_{current\ local;ir(t-1)} + \beta_3 \Delta HH\ per\ capita\ income_{ir(t-1)} + \beta_4 HH\ per\ capita\ income_{ir(t-1)} + \beta_5 SES_{child;ir(t-1)} + x_{ir(t-1)}'\gamma + \mu_r + \varepsilon_{irt} \quad \ldots(3.2.)$$

where *i* indexes individuals, *r* defines regions, *t* represents time, $x_{ir(t-1)}$ is a vector of demographic characteristics, μ_r is a time-invariant region fixed effect and ε_{irt} is a random error term. $SES_{child;it}$ is represented by recall of individuals' position on the socioeconomic ladder at age 15 in the national distribution, as well as reports of parental education.

While this OLS regression is not technically speaking a first difference regression (as it includes some variables in levels), it allows us to understand which base period indicators, along with changes in circumstances, were associated with adjustments in reports over time. Anchoring effects can be evaluated by changes in categorized subjective position in the village SES distribution, and also continuously reported income, while wave 1 *levels* of these indicators are included to understand whether individuals with higher initial incomes tended to adjust. Changes in individuals' reported

subjective well-being are introduced as an alternative source of anchoring. These questions asked individuals to rank their satisfaction with life on a scale from 1 to 10. Furthermore, the first wave's value of the dependent variable is included in levels in order to understand whether mean reversion in reporting exists, as is found in actual income data (Lechtenfeld & Zoch, 2014) and most other phenomenon reported over time (Tversky & Kahneman, 1974).

However, these regressions do not adequately control for time-invariant unobserved factors (such as ability to recall or inherent personality traits), as they only include district fixed effects. Therefore *individual* fixed effects regressions are also conducted, including person-level intercepts in the specification. The disadvantage of this approach is that many of the time-invariant covariates included in regression (3.2.) above cannot be included.

$$SES_{child;it} = \beta_0 + \beta_1 SES_{current\ local;it} + \beta_2 HH\ per\ capita\ income_{it} + x'_{irt}\gamma + \mu_i + \lambda_t + \varepsilon_{it} \quad ..(3.3)$$

with each of the indices as before, μ_i is an individual fixed effect and λ_t is a time fixed effect.

With two waves of data, coefficients from fixed effects regressions are identical to first differenced regressions (Wooldridge, 2010). As a result, it is possible to interpret all coefficients in regression (3.3.) as marginal changes in retrospective reports for a change in the covariates over the two year period, despite all variables being specified in levels. Separate analyses are conducted by gender, age cohorts and race groups, because these covariates do not vary across time. The greater length of recall for the oldest groups is investigated by differentiating results by age cohorts. Additionally, age differences may also be representative of generational differences, whereby older individuals may have a different likelihood to maintain strong family ties. The result would be better recall on parental education. Race analyses are conducted to capture a multitude of differences resulting from former separate development policies. Controls for education are introduced, as other studies have shown that the better educated are more consistent in their recall over time (Haas, 2007).

In addition, fifteen year olds in wave 1 are also analysed separately. This group is of particular interest, because retrospective data about circumstances at age 15 is enumerated in wave 2, only two years after the reference period. This additional analysis sheds light on whether length of recall is an important feature of retrospective reporting.

3.4. Results

3.4.1. Descriptive analysis

This section assesses the quality of two types of recall data on childhood circumstances. We weigh up the benefits of using objective measures of parental education (with lower variability but also lower response rates) against subjective measures of childhood SES (with higher variability and higher

response). We would expect objective assessments of parental education to be measured more consistently by respondents over time. However, the context of South Africa, where nuclear families are often fragmented, suggests that this information is also more likely to go unreported in a survey interview. In contrast, individuals may be more willing to rank their childhood socioeconomic status subjectively, as this information relates to their own experiences rather than those of their parents. However, because the measures are subjective, there may be more variability across time, and in particular, these retrospective reports may be influenced by changes in current circumstances.

Before analysing changes in retrospective reports, this section first investigates the willingness of respondents to report retrospectively. In particular, we compare whether objective measures (parental education) or subjective indicators (socioeconomic ladders) have better response rates.

Table 3.1 provides an overview of the sample of adults that appeared in both waves of the NIDS panel. Overall, nearly two-thirds of respondents rated their childhood socioeconomic status in both rounds of the survey. In contrast, parental education is particularly poorly reported, with response rates less than half that of the subjective measure. Overall, response rates are higher for mother's education than father's education. However, although this difference is statistically significant for most demographic classifications, it is not as large as we might expect given the nature of family formation in South Africa.

Table 3.1 Proportions of adult respondents in balanced panel that reported various recall measures

	Childhood SES		Mother's Education		Father's Education		N
Overall	0.655	(0.004)	0.291	(0.004)	0.277	(0.004)***	15113
Female	0.861	(0.004)	0.405	(0.006)	0.373	(0.006)***	7135
Male	0.471	(0.006)	0.189	(0.004)	0.191	(0.004)	7978
Black	0.693	(0.004)	0.323	(0.004)	0.311	(0.004)***	11689
Coloured	0.602	(0.010)	0.172	(0.008)	0.142	(0.007)***	2231
Indian	0.485	(0.032)	0.154	(0.023)	0.158	(0.024)	241
White	0.360	(0.016)	0.206	(0.013)	0.206	(0.013)	952
16-30 years	0.632	(0.006)	0.144	(0.005)	0.160	(0.005)***	5803
30-45 years	0.640	(0.008)	0.321	(0.007)	0.294	(0.007)***	4013
45-60 years	0.683	(0.008)	0.412	(0.009)	0.377	(0.009)***	3152
60+ years	0.702	(0.010)	0.457	(0.011)	0.416	(0.011)***	2113
No school	0.373	(0.008)	0.273	(0.008)	0.262	(0.008)**	3405
Primary school	0.769	(0.008)	0.401	(0.009)	0.368	(0.009)***	3026
Incomplete Secondary	0.742	(0.006)	0.246	(0.006)	0.243	(0.006)	5144
Matric	0.696	(0.010)	0.241	(0.009)	0.217	(0.009)**	2206
Post-secondary	0.707	(0.013)	0.343	(0.013)	0.337	(0.013)	1315

NOTES: Own calculations from NIDS waves 1 and 2. Figures represent the proportion of adults aged 16 years and older in the balanced panel who reported the indicator in both waves of the data, with standard errors in parentheses. Two-sided T-test of differences between response rates on mother's and father's education are significant at ***1% level **5% level *10% level. Attriters were excluded from the analysis.

Males are significantly less likely to provide retrospective reports of all types compared to females. Whites and Indians are also less likely to respond compared to the less affluent black population. This is also true for parental education, for which one would expect black respondents to have lower response, as they are more likely to live without parents[21]. Response rates increase with age for all measures, though differences between young and old are starker for the objective parental education responses. Response rates increase dramatically when adults have some education, rather than no schooling. This is the case particularly for the subjective assessment of childhood SES, suggesting that the ladder question was more easily understood by individuals with some education[22]. Response rates for parental education are also considerably higher among adults with some, rather than no, education. A possible explanation is that individuals who have some school experience are also more likely to have a better reference for what the educational attainment categories in the survey question represent. However, it is not clear why response rates for parental education then fall among those with incomplete or complete secondary education. The quality of the data is therefore compromised by highly selective response patterns. However, these patterns also suggest that adults may have been less willing to guess their parents' education than to provide a subjective assessment, so that the reports of the former may be of higher quality among respondents. In the absence of longitudinal life course data, recall data may therefore not be able to reflectively estimate the reach of childhood for a nationally representative adult population.

Selective response for objective questions indicates that those who were uncertain about parental education declined to answer, while it is possible that those who did offer this information were more certain about the true value. In contrast, subjective questions elicit high quantity responses, though they are potentially less accurate. The rest of this section compares the consistency of reports of these two measures over time, conditional on having responded to the retrospective questions in both surveys. Table 3.2 presents a set of concordance indices, both for the overall sample and for various demographic groupings. Proportions of consistent responses for individuals across time, as well as kappa measures of concordance are presented for the relative childhood socioeconomic status measure (in both its original form, and with a coarser categorization to eliminate the effects of fine marginal changes). The same analysis is conducted on parental education to assess whether consistency is indeed higher amongst those who responded due to the objective nature of the question. In addition, 95 percent confidence intervals for the kappa measures are included to analyse significant differences of concordance across groups and indicators.

[21] It is potentially true, however, that the more affluent white and Indian populations are concerned with the privacy of their information, and tend not to report these recall measures

[22] Again, a potential explanation is that more affluent individuals wish to safeguard the privacy of their information, though this is not verified.

Only 34 per cent of respondents recall the subjective ranking of childhood socioeconomic status on six rungs identically across the two-year span, while more than double that percentage provide consistent reports for parental education measures in five categories[23]. Despite better response rates for the subjective measure, it was answered less reliably. Furthermore, kappa indices are close to zero, suggesting that recall is highly erratic over time. Low concordance for this type of measure agrees with existing evidence, although in previous research different individuals were asked to provide recall information (Ward, 2011; Straughen et al. 2013). This study shows that it is not the unit of analysis that drives this result. Given that comparisons here are made for the *same* individual, our findings indicate that people are less consistent when answering subjective compared to objective questions.

Moving to the second column, where reports of childhood SES are grouped into only three larger categories, responses are more stable over time, with about double the percentage being identical across waves (62.5 per cent). Hence, many changes in reports across two waves are due to small adjustments to rankings, moving from one rung to an adjacent rung. However, the kappa measure is still decidedly modest at 0.079, with a 95% confidence interval that overlaps with that of the kappa measure for the finely categorized subjective appraisal. Hence, it is difficult to conclude that the relatively high concordance for the broad measure is a sign of consistency[24]. Additionally, even the three rung subjective measure is not as concordantly reported as the 5 category objective parental education measures. This is particularly visible in the kappa measures, which are low for both of the SES categorizations, but much higher for the recall of parental education. While the nature of these measures is very different, the descriptive evidence clearly indicates that parental education is more consistently reported than childhood SES ladders, despite poorer response rates.

Females (who have substantially higher response rates) also provide marginally more consistent responses to all measures than males, although kappa measures show that the difference is not statistically significant. Blacks, based on kappa confidence intervals, provide statistically significantly fewer concordant answers compared to coloureds on all measures. While percentages of consistent answers are somewhat lower for whites and Indians, they are not distinguishable from the other races due to imprecision in kappa statistics.

The unschooled had the lowest response rates in the education distribution, but offered the highest percentages of consistent answers across waves on all indicators. For the subjective measures,

[23] Comparisons *across* objective and subjective measures are, however, not clear cut, as concordance is also dependent on the number of categories by which each is measured, and the thresholds and distribution of the underlying latent variable that determine the categories. One can expect higher concordance when fewer categories are used, and also where there are high concentrations of respondents within particular bins.

[24] Appendix D documents transitions in reporting of childhood SES. Many individuals who initially indicated being in the upper tail of the SES distribution at childhood adjusted reports downward by more than one step in wave 2. Those at the bottom largely changed reports by one step, but almost equally upwards and downwards. Consequently, grouping adjacent steps in other combinations is unlikely to improve concordance substantially.

however, kappa confidence intervals overlap for each of the respondent education levels, so that no education effect appears to be present. For reports on parental education, kappa statistics increase significantly among respondents with more education. Comparisons of kappa confidence intervals for the recall of mother's and father's education suggest that the former was more reliably reported amongst the less educated. This pattern concurs with greater levels of absence of fathers in childhood compared to mothers. However, the difference does not arise for better-educated groups. Hence, while education generally appears to minimise recall error of objective measures (as also found by Haas (2007)) it also reduces the inconsistency of recalling father's relative to mother's education.

Given the short time (relative to time since childhood) between the evaluations, it is also unlikely that other errors such as length of recall are contaminating results. Concordance in SES rungs tends to be insensitive to respondent age, with small differences in percentages and insignificant differences in kappa statistics. Recall of parental education does, however, depend on age: kappa confidence intervals overlap only for groups aged between 30 and 60. Individuals under the age of 30 recall their parents' education with least consistency, while those above 60 show high levels of concordance. Response rates reflected the same pattern. As a result, length of recall since childhood is not problematic; rather, other potential reasons may hold. Older respondents may be part of a generation that has more integrated family links, or have had the time to re-establish these links.

A slightly different test is implemented in Table 3.3, by comparing the agreement between responses only for individuals who were 15 years old in wave 1. This limits the length of recall from the reference period to a maximum of two years. Confidence intervals for this group's kappa scores on the childhood SES measure include zero, so that the group closest to the reference period changed their assessment dramatically in wave 2, despite having experienced the recalled event most recently. This evidence suggests that substantial updating occurs in a short space of time. These changes are not likely to be the function of recall time. Parental education was reported with higher levels of consistency, although concordance is lower than for the population as a whole.

Overall then, there is considerably more variability and less concordance in replicate subjective reports than in replicate reports on parental education. Moreover, the comparison of the subjective and objective measures suggests a trade-off between response rates and consistency. High rates of subjective answers are linked with lower concordance. In contrast, low reporting rates on parental education are accompanied by more consistent data, suggesting that only the most certain individuals responded. Alternatively, because education is typically more persistent over the lifetime (as opposed to variability of current SES), recall may also be better. In the final part of this section, we explore descriptively whether changes in current assessments are related to changes in reports of each of these indicators.

Table 3.2 Proportions of respondents with identical responses over time, with kappa statistics and confidence intervals

	Childhood SES[a]	Grouped Childhood SES[b]	Mother's Education[c]	Father's Education[c]
Overall	0.340	0.625	0.726	0.727
Kappa	0.047	0.079	0.497	0.471
95% CI	(0.035 ; 0.060)	(0.060 ; 0.097)	(0.478 ; 0.517)	(0.451 ; 0.491)
Female	0.337	0.621	0.734	0.743
Kappa	0.043	0.069	0.497	0.476
95% CI	(0.027 ; 0.059)	(0.045 ; 0.092)	(0.472 ; 0.522)	(0.450 ; 0.501)
Male	0.345	0.633	0.710	0.698
Kappa	0.053	0.096	0.497	0.460
95% CI	(0.033 ; 0.074)	(0.066 ; 0.126)	(0.463 ; 0.530)	(0.428 ; 0.492)
Black	0.337	0.628	0.724	0.726
Kappa	0.030	0.036	0.438	0.393
95% CI	(0.016 ; 0.044)	(0.016 ; 0.057)	(0.415 ; 0.460)	(0.371 ; 0.416)
Coloured	0.364	0.640	0.766	0.726
Kappa	0.077	0.130	0.640	0.589
95% CI	(0.043 ; 0.111)	(0.081 ; 0.180)	(0.571 ; 0.708)	(0.518 ; 0.660)
Indian	0.308	0.530	0.649	0.658
Kappa	0.066	0.118	0.467	0.522
95% CI	(-0.037 ; 0.169)	(-0.045 ; 0.281)	(0.259 ; 0.674)	(0.332 ; 0.711)
White	0.335	0.545	0.694	0.745
Kappa	0.075	0.119	0.472	0.571
95% CI	(0.015 ; 0.135)	(0.028 ; 0.211)	(0.361 ; 0.584)	(0.464 ; 0.678)
16-30 years	0.329	0.613	0.533	0.551
Kappa	0.033	0.081	0.359	0.355
95% CI	(0.013 ; 0.054)	(0.051 ; 0.112)	(0.319 ; 0.399)	(0.316 ; 0.393)
31-45 years	0.346	0.625	0.702	0.713
Kappa	0.053	0.069	0.502	0.489
95% CI	(0.028 ; 0.077)	(0.033 ; 0.106)	(0.465 ; 0.539)	(0.451 ; 0.526)
46-60 years	0.356	0.635	0.764	0.771
Kappa	0.062	0.083	0.474	0.470
95% CI	(0.035 ; 0.089)	(0.045 ; 0.122)	(0.435 ; 0.512)	(0.431 ; 0.509)
60+ years	0.340	0.643	0.873	0.872
Kappa	0.030	0.075	0.522	0.501
95% CI	(-0.003 ; 0.062)	(0.027 ; 0.123)	(0.478 ; 0.566)	(0.455 ; 0.548)
No school	0.398	0.711	0.954	0.939
Kappa	0.045	0.035	0.251	0.134
95% CI	(0.007 ; 0.082)	(-0.015 ; 0.085)	(0.195 ; 0.308)	(0.082 ; 0.185)
Primary school	0.360	0.665	0.783	0.780
Kappa	0.018	-0.001	0.357	0.258
95% CI	(-0.01 ; 0.045)	(-0.039 ; 0.037)	(0.312 ; 0.402)	(0.212 ; 0.303)
Incomplete Secondary	0.322	0.609	0.612	0.617
Kappa	0.018	0.038	0.400	0.369
95% CI	(-0.003 ; 0.038)	(0.009 ; 0.068)	(0.363 ; 0.436)	(0.333 ; 0.405)
Matric	0.321	0.588	0.588	0.598
Kappa	0.044	0.117	0.438	0.449
95% CI	(0.013 ; 0.074)	(0.071 ; 0.164)	(0.388 ; 0.488)	(0.396 ; 0.501)
Post-secondary	0.314	0.535	0.583	0.612
Kappa	0.063	0.109	0.441	0.479
95% CI	(0.026 ; 0.100)	(0.051 ; 0.168)	(0.388 ; 0.494)	(0.425 ; 0.533)
N	9899	9899	4396	4181

NOTES: Own calculations from NIDS wave 1 and 2. The first row of each cell gives the proportion of identical reports for individuals across waves; the second row presents unweighted kappa statistics, while the third row is the 95% confidence interval of the kappa statistic. All statistics are calculated for sociodemographic characteristics based on wave 2 values. The sample includes all adults older than 15. [a]Original measure in the data, with a category for each of six rungs. [b]Rungs are grouped into three categories ("low", "middle" and "high"). [c]Grouped into "No Schooling" "Primary Schooling" "Incomplete Secondary Schooling" "Matric" "Post-Secondary Schooling"

Table 3.3 Proportions of 15-year olds in wave 1 with identical responses over time, with kappa statistics and confidence intervals

	Childhood SES[a]	Grouped Ch SES[b]	Mother's Education[c]	Father's Education[c]
Overall Prop	0.309	0.593	0.604	0.594
Kappa	0.014	0.065	0.464	0.434
95% CI	(-0.051 ; 0.078)	(-0.032 ; 0.162)	(0.302 ; 0.627)	(0.296 ; 0.571)
Female Prop	0.319	0.612	0.593	0.571
Kappa	0.021	0.052	0.454	0.43
95% CI	(-0.068 ; 0.111)	(-0.081 ; 0.186)	(0.252 ; 0.656)	(0.261 ; 0.598)
Male Prop	0.298	0.573	0.619	0.63
Kappa	-0.001	0.071	0.473	0.397
95% CI	(-0.094 ; 0.093)	(-0.070 ; 0.211)	(0.226 ; 0.72)	(0.152 ; 0.643)
N	359	359	48	69

NOTES: Own calculations from NIDS wave 1 and 2. Only individuals that were aged 15 in wave 1 were included. The first row of each cell gives the proportion of identical reports for individuals across waves; the second row presents unweighted kappa statistics, while the third row is the 95% confidence interval of the kappa statistic. All statistics are calculated for sociodemographic characteristics based on wave 2 values. The sample includes all adults older than 15. [a]Original measure in the data, with a category for each of six rungs. [b]Rungs are grouped into three categories ("low", "middle" and "high"). [c]Grouped into "No Schooling" "Primary Schooling" "Incomplete Secondary Schooling" "Matric" "Post-Secondary Schooling"

Table 3.4 explores the relationship between changes in retrospective measures and changes in individuals' perceptions of where they rank in the current village income distribution. The first row repeats the information in Table 3.2, indicating that overall, for the six rung (or grouped three rung) childhood SES measure, 34.1 per cent (62.6 percent) of assessments remain the same. An additional 27.2 per cent (12.7 per cent) of changes in childhood reports move in the same direction as changes in rankings of current income in the village distribution by the finer (coarser) classification. Movements are dominated by concordant upward changes in current and childhood rankings, at 15.8 per cent (8.1 per cent), with a slightly smaller percentage (11.3 per cent or 4.6 per cent) exhibiting concordant downward adjustments. An additional 16.4 per cent (7.6 per cent) of respondents report discordantly, with their changes in childhood and current rankings moving in opposite directions across time. A remaining 22.3 per cent (19.8 per cent) of changes in childhood rankings occur despite no adjustment to individuals' reports of their current relative circumstances. Where there is a relationship between changes in current perceptions and changes in recall of childhood SES, therefore, the modal response is a movement in the same direction. In other words, many respondents do adjust their judgment of the past in line with their perceptions of the present.

A similar analysis is also shown for the recall of parental education. It this case, discordant changes outnumber concordant changes, suggesting that individuals do not anchor reports of parental education on their perceived current position in the income distribution. However, this does not preclude the possibility

that these reports are anchored on another indicator. But given that parental education requires the recall of an objective quantity, it is likely that anchoring errors are much smaller than for subjective measures.

Table 3.4 Changes in reports on retrospective and current relative economic position

Change in direction of reports		Proportion of respondents who reported retrospective			
		Childhood SES		Parental Education	
Past	Current	Original[a]	Grouped[b]	Father's[c]	Mother's[c]
Unchanged		0.341	0.626	0.728	0.727
↑	↑	0.158	0.081	0.033	0.033
↓	↓	0.113	0.046	0.051	0.049
↑	↓	0.080	0.034	0.032	0.033
↓	↑	0.084	0.042	0.067	0.068
↑	Unchanged	0.120	0.088	0.036	0.035
↓	Unchanged	0.103	0.084	0.053	0.053
Total concordant changes		0.272	0.127	0.084	0.082
Total discordant changes		0.164	0.076	0.099	0.102

NOTES: Own calculations from NIDS wave 1 and 2. [a]Original measure in the data, with a category for each of six rungs. [b]Rungs are grouped into three categories ("low", "middle" and "high"). [c]Grouped into "No Schooling" "Primary Schooling" "Incomplete Secondary Schooling" "Matric" "Post-Secondary Schooling"

3.4.2. Regression estimates

We now turn to estimating the correlates of changes in retrospective reports using first differenced OLS (3.2.) and fixed effects regressions (3.3). This sheds light, for instance, on whether changes in perceptions of current income rankings really concordant with retrospective reports, even once conditioning on other factors such as reported income, education and person-specific unobservables. Table 3.5 presents regressions of changes in reports of retrospective measures across waves. Perceived relative current socioeconomic status is clearly associated with changes in assessments of childhood SES. Individuals who ranked their households on a relatively high income step in wave 1 were more likely to raise their rankings of childhood SES. Moreover, individuals who changed their assessment of current economic status relative to the village tended to change their ratings of childhood in the same direction. Levels and changes in *actual* household income had similar relationships with adjustments in retrospective SES reports. Individuals also changed their childhood SES reports concordantly with adjustments to their subjective well-being. In contrast, changing reports on parental education are not associated with similar movements in subjective rankings of current income or life satisfaction. This confirms conclusions from Table 3.4 that recall of parental education does not appear to be anchored by changes in subjective reports of current circumstances. However, individuals who lived in households with higher levels of reported income and with increases in this income tended to raise their reports of their parents' education.

Each of the specifications also controls for the level from which the respective dependent variables changed in wave 1, in order to understand whether regression to the mean occurred in reporting (Tversky & Kahneman, 1974). This was the case for each indicator, whereby those with lower prior reports tended to raise them.

While the descriptive statistics presented in Table 3.2 highlight that females were more consistent than males in their reports of childhood SES, the regressions in Table 3.5 show that among those who did change their reports, females raised their assessments more than males over time. No systematic gender differences occur for assessments of parental education. Whites also raised their reports on all indicators by the highest magnitudes. Age has no influence on changes in recall of childhood SES, confirming again that length of recall is not a factor influencing the consistency of these reports. However, older individuals tended to lower their reports of parental education over time. As education increased, individuals tended to adapt answers to all questions upwards across waves, pointing to a potential learning effect. It is, however, possible that parental education measures are anchored on own education rather than on current socioeconomic status.

Table 3.5 Correlates of changes in retrospective reports, differenced OLS regressions

OLS	ΔChildhood SES[a]	ΔMother's education[b]	ΔFather's education[b]
Current village income step$_{t-1}$	0.242	0.001	0.004
	(0.016)***	(0.071)	(0.074)
ΔCurrent village income step	0.195	-0.068	-0.077
	(0.011)***	(0.047)	(0.048)
log(real pc HH income)$_{t-1}$	0.117	0.183	0.244
	(0.014)***	(0.063)***	(0.065)***
Δlog(real pc HH income)	0.070	0.199	0.251
	(0.013)***	(0.058)***	(0.060)***
Childhood SES step$_{t-1}$	-0.978		
	(0.011)***		
Mother's education$_{t-1}$		-0.525	
		(0.015)***	
Father's education$_{t-1}$			-0.537
			(0.015)***
ΔSubjective well-being	0.024	-0.029	-0.025
	(0.004)***	(0.016)*	(0.017)
Female	0.045	-0.003	0.076
	(0.021)**	(0.098)	(0.100)
Coloured	0.174	0.141	1.306
	(0.052)***	(0.270)	(0.308)***
Indian	0.004	1.145	1.690
	(0.096)	(0.512)**	(0.497)***
White	0.496	2.968	3.180
	(0.065)***	(0.274)***	(0.289)***
Age	0.000	-0.027	-0.030
	(0.001)	(0.004)***	(0.004)***
Primary education	0.040	-0.281	-0.350
	(0.038)	(0.144)*	(0.150)**
Incomplete secondary	0.153	0.478	0.341
	(0.041)***	(0.166)***	(0.173)**
Matric	0.197	0.911	0.612
	(0.048)***	(0.212)***	(0.220)***
Post-secondary	0.231	1.035	1.060
	(0.053)***	(0.229)***	(0.236)***
Constant	0.091	1.310	0.130
	(0.144)	(0.653)**	(0.689)
District council FE	Y	Y	Y
R-squared	0.535	0.292	0.309
N	7882	3454	3280

NOTES: * p<0.1, ** p<0.05, *** p<0.01. Own calculations from National Income Dynamics Study, waves 1 and 2. Standard errors in parentheses. [a]Original measure in the data, with a category for each of six rungs, differenced over time. [b]Measured in years of education, differenced over time.

The remaining three tables in this section (Table 3.6 to Table 3.8) present fixed effects regressions investigating the factors that influence the changes in adults' retrospective reports. In contrast to regressions reported in Table 3.5, unobservable traits are now also accounted for. Analyses are conducted

for various sub-samples to understand whether anchoring varies by group[25]. Specifications include current relative position in the village income distribution, subjective well-being and also reported per capita household income, to measure whether anchoring occurs based on changes in real economic circumstances. As noted before, even though these regressions are specified in levels, they can be interpreted in exactly the same way as first difference regressions.

For most sub-samples, changes in individuals' subjective ranking of current income moves concordantly with changes in rankings of adolescent socioeconomic status in Table 3.6. The adjustment is about 0.145 of a childhood step for every category jump in perceived current status. This estimate is only slightly lower than that in Table 3.5, suggesting that fixed effects do not influence conclusions to a large degree. The anchoring effect is similar across genders. However, the youth (16 to 29 years) adjust their view of the past by about one fifth of a step for every change in current ranking, while those older than 60 do so by roughly half the amount. It is therefore apparent, together with the evidence in Table 3.3, that length of recall since childhood is not influential in changing perceptions of the past; rather, the youth are more prone to update their assessments over time, with a potential learning effect occurring among older individuals.

Reported current household income improvements are significantly associated with a more optimistic view of the past for the overall population in Table 3.6, though the effect is dominated by females, those aged 30-59 and blacks. While not shown, these effects disappear once a more aggregated classification of childhood SES is used as a dependent variable, suggesting that changes in reported income only result in small shifts in evaluating the past[26]. Additionally, changes in subjective rankings of well-being are also significantly concordant with adjustments to childhood SES rankings. It is therefore evident that changes to current *perceptions* have a more important role in explaining how people change their view of the past, with a smaller emphasis on changes in manifest household economic circumstances. Individuals therefore first form perceptions about their current socioeconomic status and subjective well-being, after which they adjust from this anchor to arrive at their assessment of past socioeconomic status.

Education is introduced as a control in most regressions. It only plays a minor role in individuals' changing assessments of adolescent socioeconomic status, with many insignificant coefficients across most specifications in Table 3.6. Estimates in Table 3.5 suggest that education does, however, have a potential influence on changes in recall. However, the small time variations in the data cannot adequately identify these relationships in fixed effects regressions.

[25] This approach is followed, primarily because these groups are persistent across time, so that the classifications cannot be introduced as control variables in fixed effects estimates.
[26] All other results are robust to this change in the dependent variable.

Table 3.7 and Table 3.8 repeat the exercise with objective measures of parental education. Notably, changes in individuals' perceptions of their relative position in the village income distribution and changes in subjective well-being do not have the same role to play in reporting on parental education as they did for childhood SES ladders. These results are similar to estimates in Table 3.5. The exception is that females who rated their relative income one category higher reduced their estimate of parental education by about one tenth of a year. Changes in reported incomes have more of a role to play in this instance, especially for reports of paternal education. A one percentage increase in household per capita income raised estimates of father's education by close to one fifth of a year, although this effect is dominated by female respondents. The association increased from a statistically insignificant value for those aged below 30 to roughly 0.3 years for the elderly. The association between changes in reported income and changes in recall of father's education is significant, consistent with anchoring effects. This association increases among older respondents. Associations for maternal education are weaker, so that anchoring on income is less severe for this indicator. Individuals therefore appear to be more certain about maternal education than paternal education, as the reports of the former are more consistent over time. This concurs with the high levels of absent fathers in South Africa.

Additionally, reports of father's education were particularly sensitive to the increases in respondents' qualifications. Males who matriculated during the period lowered their reports of their father's education by almost two years, which also agrees with a similar pattern among those aged below 30. Males also adjusted their reports downwards if they obtained a post-secondary qualification. While the effects are much smaller compared to the group of males, blacks displayed a similar pattern. Hence, reports of paternal education are potentially anchored on own education, but especially when individuals improve their qualifications. Table 3.5 showed, however, that *levels* of own education are also influential.

The same cannot be said for learning about mother's education levels. Individuals were more likely to live with mothers during their youth and potentially recall information about them consistently. However, all education coefficients are obtained from small samples of switchers, so that these results are not conclusive. Furthermore, ratings of own education could have been inconsistent, contributing to a concordant increase in individuals' assessments of themselves and their parents.

In summary, individuals anchor their perceptions of past SES on perceptions of current household circumstances, but to some degree also on manifest changes in income. This type of anchoring is not as prevalent when objective measures of childhood status are used. However, some learning does appear to occur over time in assessing paternal education, particularly if males obtain more education themselves.

Table 3.6 Fixed effects regressions for changes in reports of childhood SES

	Dependent variable: Socioeconomic ladder rung at age 15 (6 rungs)						
Sample:	All	Female	Male	Age 16-29	Age 30-59	Age 60+	Black
Current village income step	0.145	0.142	0.148	0.176	0.132	0.080	0.142
	(0.012)***	(0.015)***	(0.019)***	(0.019)***	(0.018)***	(0.033)**	(0.012)***
log(pc HH income)	0.037	0.055	0.018	0.027	0.041	0.031	0.049
	(0.016)**	(0.021)***	(0.025)	(0.024)	(0.024)*	(0.058)	(0.017)***
Subjective well-being	0.012	0.009	0.018	0.021	-0.003	0.029	0.013
	(0.005)**	(0.006)	(0.008)**	(0.008)***	(0.007)	(0.014)**	(0.005)**
Matric	-0.011	-0.105	0.141	0.052	-0.103	1.029	-0.048
	(0.075)	(0.101)	(0.124)	(0.089)	(0.207)	(0.460)**	(0.082)
Post-secondary education	-0.034	-0.276	0.27	0.092	-0.022	0.486	-0.064
	(0.095)	(0.135)**	(0.145)*	(0.136)	(0.194)	(0.483)	(0.107)
Year = 2010	0.096	0.102	0.076	0.01	0.128	0.156	0.12
	(0.016)***	(0.020)***	(0.026)***	(0.027)	(0.024)***	(0.045)***	(0.017)***
Constant	1.328	1.281	1.356	1.353	1.372	1.252	1.218
	(0.108)***	(0.141)***	(0.175)***	(0.158)***	(0.175)***	(0.406)***	(0.113)***
Individual FE	Y	Y	Y	Y	Y	Y	Y
R-squared	0.030	0.029	0.032	0.040	0.029	0.029	0.036
N	26838	15749	11081	11346	12087	3405	21440

NOTES: * p<0.1, ** p<0.05, *** p<0.01. Own calculations from National Income Dynamics Study, 2008 and 2010 waves. All variables are measured in levels, though interpretation is the same as a first difference regression. Standard errors in parentheses.

Table 3.7 Fixed effects regressions for changes in reports of maternal education

	Dependent variable: Mother's education in years						
Sample	All	Female	Male	Age 16-29	Age 30-59	Age 60+	Black
Current village income step	-0.046	-0.105	0.032	0.037	-0.083	0.018	-0.03
	(0.041)	(0.049)**	(0.072)	(0.12)	(0.054)	(0.054)	(0.043)
log(pc HH income)	0.157	0.133	0.175	0.252	0.123	0.052	0.152
	(0.057)***	(0.068)*	(0.101)*	(0.145)*	(0.073)*	(0.112)	(0.061)**
Subjective well-being	-0.005	0.001	-0.005	-0.075	0.004	-0.003	-0.008
	(0.017)	(0.021)	(0.03)	(0.053)	(0.022)	(0.024)	(0.019)
Matric	-0.414	-0.509	-0.51	-0.919	0.138	2.572	-0.423
	(0.394)	(0.488)	(0.661)	(0.67)	(0.807)	(0.937)***	(0.444)
Post-secondary education	0.108	-0.061	-0.065	-1.12	1.145	3.985	-0.006
	(0.432)	(0.592)	(0.665)	(0.981)	(0.726)	(1.051)***	(0.497)
Year = 2010	-0.418	-0.388	-0.519	-0.816	-0.416	-0.199	-0.392
	(0.055)***	(0.065)***	(0.100)***	(0.177)***	(0.070)***	(0.077)***	(0.061)***
Constant	2.604	2.677	2.696	4.73	2.158	0.807	2.063
	(0.394)***	(0.466)***	(0.704)***	(0.971)***	(0.531)***	(0.759)	(0.406)***
Individual FE	Y	Y	Y	Y	Y	Y	Y
R-squared	0.021	0.023	0.027	0.045	0.025	0.033	0.018
N	14887	9129	5757	4180	8073	2634	12239

NOTES: * p<0.1, ** p<0.05, *** p<0.01. Own calculations from National Income Dynamics Study, 2008 and 2010 waves. All variables are measured in levels, though interpretation is the same as a first difference regression. Standard errors in parentheses.

Table 3.8 Fixed effects regressions for changes in reports of paternal education

	Dependent variable: Father's education in years						
Sample	All	Female	Male	Age 16-29	Age 30-59	Age 60+	Black
Current village income step	-0.045	-0.133	0.061	-0.021	-0.035	-0.047	-0.033
	(0.042)	(0.054)**	(0.067)	(0.117)	(0.056)	(0.063)	(0.045)
log(pc HH income)	0.166	0.181	0.081	0.128	0.201	0.211	0.169
	(0.060)***	(0.078)**	(0.093)	(0.148)	(0.078)**	(0.111)*	(0.065)***
Subjective well-being	-0.002	-0.023	0.034	-0.047	-0.007	0.016	-0.011
	(0.018)	(0.022)	(0.029)	(0.048)	(0.024)	(0.027)	(0.02)
Matric	-0.854	-0.431	-1.542	-1.718	-0.272	1.049	-0.918
	(0.348)**	(0.452)	(0.565)***	(0.615)***	(0.719)	(1.096)	(0.381)**
Post-secondary education	-0.665	-0.204	-1.061	-1.573	-0.348	1.525	-0.915
	(0.394)*	(0.527)	(0.602)*	(0.893)*	(0.638)	(1.008)	(0.457)**
Year = 2010	-0.436	-0.39	-0.617	-0.760	-0.438	-0.256	-0.443
	(0.059)***	(0.073)***	(0.097)***	(0.177)***	(0.075)***	(0.088)***	(0.066)***
Constant	2.554	2.453	3.216	5.17	1.724	0.000	1.946
	(0.413)***	(0.525)***	(0.647)***	(0.977)***	(0.561)***	(0.775)	(0.429)***
Individual FE	Y	Y	Y	Y	Y	Y	Y
R-squared	0.022	0.021	0.048	0.045	0.026	0.025	0.022
N	16307	9758	6546	5147	8489	2671	13591

NOTES: * p<0.1, ** p<0.05, *** p<0.01. Own calculations from National Income Dynamics Study, 2008 and 2010 waves. All variables are measured in levels, though interpretation is the same as a first difference regression. Standard errors in parentheses.

3.4.3. Implications for using retrospective data

Recall of parental education is less prone to anchoring than subjective measures. However, response rates in South Africa are low due to parental absence. Highly selective samples compromise the use of parental education indicators in further analysis. The alternative would be to use subjective rankings of childhood SES. In light of the analysis above it is clear, however, that individuals base their subjective perception of the past on current circumstances and events. As a result, they are contaminated measures that are enumerated with error (potentially in both rounds of data). If measurement errors are uncorrelated, it is warranted to use another enumeration of the same variable as an instrument in analysing causal effects – for instance, to establish the childhood reach on adult outcomes in the labour market. However, if persistent misperceptions contaminate data, IV estimators only partially solve for these problems, as illustrated using Australian health rankings (Crossley & Kennedy, 2002). The IV reduces attenuation bias, though does not solve it.

Anchoring in perception data can have even more severe implications for IV analysis, beyond correlated measurement errors. This presents an issue that is not commonly raised in the literature. The reason for this is that the anchoring effect may *result* from changes in the outcome variable (such as employment or health), raising the level of measurement error in the assessment of the past. As a result, reverse causality arises, further attenuating coefficient estimates. Repeated measures of childhood SES will suffer from the same problem, so that this approach does not lend itself to causal analysis of childhood reach. The instrument will be endogenous to the outcome variable, as anchoring represents a common measurement error to both enumerations. These issues may be of particular relevance in a developing country, where life course studies are less prevalent *and* response rates and consistency of recall questions are poorer. Additionally, this type of measurement error compromises the use of childhood measures in household bargaining models (Browning & Bonke, 2009; Browning & Lechene, 2003).

3.5. Conclusions

Retrospective measures fill gaps in information where evaluated life histories are unavailable, and are particularly important in developing countries. In South Africa, knowledge about life course persistence could contribute to a broader literature on the relative lack of mobility within this society. However, the reliability of such indicators becomes questionable when their scale is relative and when subjective reports are sensitive to anchoring effects. This study has illustrated that subjective reports of childhood SES elicit higher response rates than objective measures of paternal and maternal education. Nevertheless, response is highly selective. Additionally, the responses that were collected are sensitive to changes in both perceived and manifest current economic circumstances, with the former being more important. Objective measures, however, do not necessarily offer a better

alternative, as advocated by others (Haas, 2007). While maternal education is rated with high consistency and with minor anchoring effects, this indicator has distinctly poor response rates. Paternal education has even worse response. Furthermore, length of recall and higher educational attainment matter somewhat for changes in reports of father's education. This is reflective of a society that is dominated by single parent households, and where (objective) information on respondents' fathers may not be trustworthy. A trade-off between response rates and consistency of reporting between different evaluations of the past therefore exists, leaving substantially contaminated data with which to assess the reach of childhood. Without detailed knowledge of the nature of measurement error that arises in recall data, its validity in further analysis is compromised.

Researchers should be aware of the fact that, especially where severe anchoring occurs, instrumenting for childhood socioeconomic status with another report of the same indicator is not necessarily an improvement on the situation. Potential anchoring on the adult outcome variable (such as employment status) may influence the dependent variable (the rating of childhood SES), as well as the instrument, so that additional reverse causality is introduced by such an approach. Instruments are unlikely to be exogenous, yielding inconsistent estimates of childhood reach. Above the possibility of correlated measurement errors over time, the effects of anchoring in causal estimation have not been studied in great detail. This study alerts researchers to the potential pitfalls of estimating life course mobility with retrospective data, and prompts for additional investigation in how to use it credibly.

Chapter 4

Separating temporary from permanent generational change in labour force participation – heterogeneous age-period-cohort models for South Africa

4.1. Introduction

Substantial increases in labour force participation pre-date the transition to democracy in South Africa, but have contributed to a rising unemployment rate in the post-apartheid period (Banerjee et al., 2008). Should the pressure to provide jobs for a growing labour force not be alleviated, this can result in a permanent high unemployment equilibrium. It is therefore helpful to understand whether various changes in labour force participation are likely to persist, or whether their effects may reverse or dissipate over time. Participation increases can be classified into potential long-term movements – such as the feminization of the labour force (Casale & Posel, 2002) or the pull of higher wages since the deregulation of black unions in the 1970s (Lewis, 2001) – and possible short-run shifts, such as large-scale labour market entry following the implementation of a post-apartheid policy that was devised to unburden the overcrowded schooling system (Burger et al., 2014). The latter limited the ages at which learners could stay in school. While Banerjee et al. (2008) note that some of the long-run features are likely to have permanent effects on the labour market, it is not certain whether other tendencies could change behaviour over a long time horizon. The objective of this paper is to distinguish between generational change that is persistent and that which is temporary.

This objective is closely related to the literature on age-period-cohort (APC) analysis, which decomposes demographic change into life cycle, aggregate time and generation-specific movements. In particular, model identification strategies assume that generational change is permanent throughout birth cohorts' lifetimes, while life cycle trajectories represent substantial variation over time for all cohorts, and periodic fluctuations are more temporary in nature, affecting an entire economy only in certain periods. Should the rise in participation be primarily generational, new cohorts that added to the labour market stock are unlikely to change their behaviour and leave the job queue. However, if these groups behaved in this manner only for a part of their lifetimes, their participation rates may moderate as they age, shedding a part of the labour surplus. Similarly, temporary economy-wide variation may also reverse, prompting all cohorts to ease labour supply and reduce unemployment. The central question of this chapter, therefore, is which components of generational labour supply are likely to remain permanent. Might one expect participation to return to previous equilibria, thereby lowering unemployment?

Tests of these hypotheses, however, require methodological adaptations that have not been commonly applied in either the labour economics or the more general APC literatures. In particular, *additive* APC models have standardly been applied in separating out generational from other time variation. The one shortcoming of such a model is that cohort (or generational) effects are implicitly assumed to be permanent (Glenn, 1976). In the current context, cohort-specific growth in participation is assumed to remain constant over these groups' life cycles. This study adds to this literature, by investigating the plausibility of cohort effects that can be heterogeneous over the life cycle. To do so, models that interact age and cohort profiles are estimated. By implication, some generations may behave distinctly for sub-periods of their lives, with the rest of their life cycles resembling those of other cohorts. In such a case, change may only be interpreted as temporary.

Interactive APC modelling presents various challenges, however. Because of a well-known identification problem in simpler additive APC models, few researchers have extended this specification to relax the assumption of permanent cohort effects. In particular, because the current year is just the sum of all groups' birth years and their current ages, APC models suffer from perfect multicollinearity. Interactive specifications with heterogeneous cohort profiles compound this problem, with each of the cross terms also linearly related to each other. This chapter presents a thorough investigation of a range of identification strategies proposed in the literature, ranging from imposing behavioural assumptions (Deaton, 1997) to the promising (but criticized) intrinsic estimator (Yang et al., 2004) to multidimensional non-parametric alternatives (Hastie & Tibshirani, 1990) and those that use maximum entropy principles (Browning et al., 2012). These estimators are evaluated as contenders for offering a base solution to the heterogeneous cohort profiles that most likely underlie economic phenomenon. Monte-Carlo simulations show that each of these approaches has its own weaknesses in both additive and interactive scenarios.

However, the very practical approach of searching for testable behavioural restrictions in additive models (McKenzie, 2006) can be eloquently expanded to delineate multi-dimensional regions of the data for which particular functional assumptions hold. This allows for the effective implementation of interactive models. Illustrations with Labour Force Survey data highlight that previous studies correctly identified cohort behaviour with traditional methods. However, the expansion to interactive methods indicates that, while a part of this cohort trend is long-run and permanent, other distinct patterns are likely to be temporary in nature. Rather, they represent a deviation in behaviour for a subset of cohorts, and only for a limited period in their life cycles. In the former scenario, policy would have to focus on long-run solutions, while in the latter case only temporary intervention may be required until this group's behaviour normalizes. Hence, this chapter provides a nuanced view of generational change, and using appropriate methodology sheds light on which components of labour force participation are more structural in nature, and which are likely to dissipate.

The rest of this chapter proceeds as follows. Section 4.2 provides selected highlights from the extensive literature on the APC problem, focusing on the array of existing identification assumptions of the additive model. It also seeks to find potential solutions to the heterogeneous model. Section 4.3 embarks on a Monte-Carlo simulation study to assess the adequacy of additive and interactive solutions on a data generating process that contains heterogeneous age and cohort profiles and cyclical time variation. Section 4.4 applies the solutions to the labour force participation scenario that was sketched above. Section 4.5 concludes.

4.2. Motivation and identification of APC models

4.2.1. Motivation for this study

The central question of this chapter is to understand the permanence of generational change in South Africa's labour force. Post-apartheid schooling policies (Burger et al., 2014), the feminization of the labour force (due to higher educational attainment and lower marriage rates) (Casale & Posel, 2002), the increase in unskilled wages since black unions were unbanned in the 1970s (Lewis, 2001) and changes in cultural norms have all contributed to the increase in labour supply. Whilst additive age-period-cohort (APC) decompositions are useful empirical tools to separate the generational component from aggregate time trends and life cycle effects, their previous implementation has ignored the potential that generational effects can be temporary so that as cohorts age, distinct behaviour might normalize in line with other groups. In such a case, the policy may therefore only need to be temporary as the distinct behavioural pattern is specific to a generation only at a particular point in their lives. Is the increase in labour force participation in South Africa a permanent feature, with younger generations likely to stay in the job queue in greater numbers even later in life? Or will increases in participation slow down as these groups age, matching the behaviour of older generations? Initial indications suggest that at least the long-run participation increases attributed to feminization are unlikely to reverse (Banerjee et al., 2008). However, it is not clear whether other causes of labour force growth will be sustained. To this end, existing work on APC analysis will be interrogated and adapted.

Additive APC decompositions have long been implemented to separate distinct, but related, temporal variation in demographic phenomena. Epidemiologists, sociologists and economists have found great utility in their description of change. These models are analogous to business cycle analysis, which separates long-run trends from periodic fluctuations in the economy. APC analysis does much the same, yet it follows micro units, and accounts explicitly for the heterogeneity in aggregate change experienced by various age groups and generations. The typical additive model decomposes an indicator of choice (y_{apc}) into age (α_a), period (π_p) and (ζ_C) cohort components (with an exhaustive set of dummy variables) using linear models as a basis:

$$y_{apc} = \mu + \alpha_c + \pi_p + \zeta_C + \varepsilon_{apc} \ldots (4.1)$$

$$\text{where } a = 1,2 \ldots, A;\ p = 1,2 \ldots, P;\ c = 1,2 \ldots, C$$

In any given period each individual of a certain age group can only belong to one year of birth. These unique cells (henceforth also referred to as *apc* cells) can be identified as the units of observation in repeated cross sections. This option is not only appealing, but also mostly valid if each of the *apc* cells contains a sufficient number of observations, usually one hundred or more (Verbeek & Nijman, 1992).

The discussion of decomposing demographic variables into these three components has primarily focussed on the inability of saturated dummy variable linear models to uniquely identify these effects. The design matrix is singular for the simple reason that *period = birth year + age*, resulting in perfect collinearity of regressors that, however, are supposed to capture distinct phenomenon. The central pursuit of this literature has therefore been credible identification, and a multitude of identifying restrictions has been proposed by researchers from a range of disciplines to arrive at estimates that additively separate the three components (Deaton & Paxon, 1994; Deaton, 1997; Yang et al., 2008; Browning et al., 2012; Yang et al., 2004; Keyes & Li, 2010; Jiang & Carriere, 2014). A selection of these identifying restrictions will be discussed below. Regardless of the (sometimes seemingly arbitrary) choice of restriction, the shapes of each of the age, period and cohort profiles tend to remain the same, with only their slopes exhibiting differences (McKenzie, 2006). Despite these (mild) consistencies, the ultimate choice of method is not a clear-cut case, especially since divergent profile slopes can emphasize one component of the decomposition in favour of another. For example, period effects may be estimated as cyclical or short-run, emphasizing cohort effects instead. Alternatively, long-run change could be estimated as a period effect, with little role for cohort change. The difference is that the former scenario is generation-specific, while the latter change affects all cohorts in the same way. Because each of the effects has a distinct interpretation, the sensitivity of estimates to their associated identification strategy is the most important consideration in this class of analysis. Each of the effects will now be discussed in turn.

Period effects have the most intuitive interpretation and represent the circumstantial ebbs and flows that affect all members of the population of interest at any point in time. For economists, this quantity can represent two types of movement. Firstly, it could denote long-run change that is not specific to age groups or cohorts. Secondly, it may represent a cyclical component of any time series, and is akin to an instantaneous movement, which may subside in the subsequent period. Period effects that underlie economic phenomenon are often determined by the phases of the overall business cycle, and track how short-run macroeconomic movements affect micro behaviour.

Age effects capture life cycle behaviour of the typical individual, regardless of the generation they belong to, or the economic circumstances within which they make their decisions. Typical life cycle patterns are explicit in most microeconomic characterisations of behaviour: the most famous is the concave earnings life cycle observed within most labour forces (Mincer, 1962), the accompanying aspirations of individuals (Easterlin, 2001) and the pattern of labour force participation that is matched by these returns. The permanent income hypothesis (Friedman, 1957) distinguishes between consumption patterns early in life (when individuals tend to dis-save, borrowing from anticipated future income) and later in life (when debts are paid off or existing precautionary savings are spent), predicting a smoother pattern of expenditure over a lifetime relative to income. Each of these life cycle patterns is central to economic theory.

Cohort patterns are easily confused with life cycle effects, as a group of different birth cohorts represent a number of age groups at any point in time. However, cohort patterns are often assumed to remain permanent over time for groups that were born in specific years or are members of particular generations. It is not difficult to imagine that certain generations have distinct spending behaviour (such as the new emerging black middle class in South Africa (Kaus, 2013)) or attitudes (such as the baby boomer generation or "generation Y"). Furthermore, policy changes often affect particular groups: regulations that are enacted when individuals enter the labour market can potentially follow them for the rest of their lives. Distinct labour market circumstances that are experienced by a specific cohort (such as a recession) can have lasting effects that are specific to that cohort throughout its life cycle. For instance, Gregg and Tominey (2005) document a long-term persistent effect on wages following a period of initial unemployment. *Additive* APC models assume the permanence of such effects, with the understanding that behaviour is determined by cultural norms and traditions from which certain generations do not deviate, and that initial conditions are slow to dissipate.

This latter assumption is, however, restrictive if cohort shocks are not permanent, and generations are able to adjust to changing economic circumstances. Glenn (1976) emphasizes that while "formative experiences" may affect specific cohorts permanently, new experiences can add to this status quo, so that both old and new "information" is implicit in cohort-related outcomes. The variation is therefore not likely to be permanent. For instance, spending patterns of today's black middle class in South Africa may change in future. Furthermore, life cycle patterns can potentially change over time, with various generations following a distinct course as they age. In all of these circumstances, age, period and cohort effects (though distinct in their interpretation) are confounded by each other. The only way to relax the assumption of permanent generational effects (or identical life cycle profiles for all generations), is to extend the APC model to be interactive.

To illustrate the need for temporary cohort effects, consider the example of South African labour force participation that has been analysed elsewhere using additive APC models (Burger & von

Fintel, 2014; Burger et al., 2014). Regardless of identification assumptions, it is clear that cohorts of blacks born after 1975 entered the labour market disproportionately, over and above other long-run generational increases. A large part of this pattern can be ascribed to the introduction of post-apartheid schooling policies, which pushed over-aged learners out of classrooms and into the job queue. This pattern manifests in large estimated cohort effects for the youngest generations, though these differences may not be sustained across these groups' lifetimes, especially at ages when these individuals would have entered the labour market even in the absence of the policy. Furthermore, schooling policies in the past are unlikely to influence retirement decisions, so that the behaviour of these cohorts later in the lifecycle may very well resemble those generations that preceded them. Such patterns cannot be observed until data is collected for these cohorts much later in life. Typically, therefore, such eventualities are rarely accounted for, due to a short series of surveys that do not span entire lifetimes.

A more accurate reflection of this situation is a differentiated age profile for younger generations, but only at the beginning of their working lives. Recent generations show greater propensity for labour market entry at earlier ages, while they may as well retire at the same age as earlier generations. Exactly these patterns are documented non-parametrically by Branson and Wittenberg (2007), although their analysis does not distinguish between aggregate period effects and those that are specific to some generations. Hence, this type of behaviour prompts the need to model interactive age effects for various cohorts, which accounts for generationally distinct age patterns, potentially also only for certain parts of the life cycle.

While a scenario of long-run change (such as feminization of the labour force) is amenable to estimation with additive APC models (where one common cohort effect is established for those born in the same year), the schooling scenario sketched here defies APC assumptions, as the (distinct) cohort effect also depends on the point in the life cycle at which these groups find themselves. Hence, abstracting from the identification problems of additive models, additional misspecification arises, as age and cohort are now also interactively intertwined.

Before discussing this central point, this paper first reviews some of the most widely used identification strategies in *additive* APC analysis. The objective is to inform identification strategies for *interactive* solutions, the ultimate goal of this research. Usually one identification restriction must be imposed, so that the other coefficients can be estimated without further problems. However, other solutions that do not depend on explicit identification restrictions have also been proposed in the literature. Because APC analyses essentially amount to a design matrix (X), where the columns represent the dummy variables from the saturated additive model) that has a rank of one less than the full column rank, $X'X$ cannot be inverted. This incomplete rank is associated with one eigenvalue that equals zero, which is symptomatic of the singularity of the design matrix. This deficiency is

equivalent to the need for at least one restriction to estimate APC models. Hence, $\widehat{\boldsymbol{\beta}}_{OLS} = (X'X)^{-1}X'\boldsymbol{y}$ is not defined and requires additional identification assumptions.

4.2.2. Extending identifying assumptions of additive models to interactive models

Age-period-cohort (APC) decompositions have featured in the literature of multiple disciplines for decades, and new proposals to solve an age-old problem are still offered at regular intervals. Despite early proclamations that this type of analysis is a "futile quest" (Glenn, 1976), with recent re-iterations labelling it more extremely as an "unholy quest" (Fienberg, 2013), attempts to perfect this model continue. It is, however, the descriptive appeal of APC estimators that lure researchers to different variants thereof, and to find the most credible solutions to interpret their findings. Hence, it is unlikely that applied scientists will abandon the approach, nor will their attempts to solve the puzzle cease. However, given the continued proliferation of APC analysis, applied researchers should be aware of the identification assumptions that determine their conclusions. More importantly, they should be aware that some solutions that have been presented as "assumption agnostic" do not in fact live up to that promise, and should not be applied without careful thought about the influence of the specific data generating process (Luo, 2013). This section reviews some of the identification strategies that have emerged in the last few decades. In each case the possibility of extending this estimator to the interactive setting is evaluated.

4.2.2.1. Coefficient restrictions based on behavioural assumptions

A priori theory can often simplify the identification of APC models. For instance, Mincerian earnings functions (Mincer, 1962) entail the specification of a quadratic in experience (or age). Following this functional form for age effects allows cohort and period effects to be flexibly specified without fear of collinearity. The problem with this approach, however, is the rigidity that this imposes on an entire life cycle, which may not hold precisely in the population. Grün (2003) has analysed South African wages in this manner. Such an approach is amenable to interactive effects, with the possibility of estimating various parametric age profiles for specific cohorts. However, the validity of the quadratic parametric assumptions is uncertain without further verification.

Alternatively, many authors choose to restrict any one coefficient to zero or to be equal to that of an adjacent group, such as forcing two time periods to have an equivalent effect. This was famously proposed by Mason et al. (1973). However, as Browning et al. (2012) illustrate, results are sensitive to such seemingly minor behavioural assumptions, and require clear motivation to be valid (such as two adjacent cohorts that really did face identical long-term common effects).

A widely-used method in economics turns to restrictions on the period effects, so that they average to zero and are orthogonal to time (Deaton & Paxon, 1994; Deaton, 1997). To do so, two time dummies

are omitted, while the others are adjusted by the following transformation and included in a standard APC regression (instead of the usual time dummies):

$$\pi_p^* = \pi_p - [(p-1)\pi_2 - (p-2)\pi_1] \quad p = 3, \dots, P \dots (4.2)$$

where *P* is the total number of periods contained in the data. Subsequently, the effects of the first two periods can be recovered as follows:

$$\pi_1^* = \pi_3^* + 2\pi_4^* + \cdots + (T-2)\pi_T^*$$

$$\pi_2^* = -2\pi_3^* - 3\pi_4^* - \cdots - (T-1)\pi_T^* \dots (4.3)$$

This orthogonality condition is steeped in the understanding of business cycles, in that short-run fluctuations do not contain a sustained upward or downward trend. The behavioural assumption underlying this strategy is that period effects average to zero over time. Long-run trends in any data are therefore relegated to cohort and age effects. For instance, growth in labour force participation may be represented by greater entry among more recent cohorts, while it is potentially true that all cohorts entered in greater numbers. In this case, the behavioural assumption imposed by this restriction would be incorrect. While this restriction is also theory-based, and likely to be true for many outcomes that are directly connected to economic activity, its value is limited if the data does not extend over a sufficient section of the business cycle (Deaton, 1997). Furthermore, if time variation is truly long-run and acyclical, this estimator could perform poorly. It is also not certain how this restriction can be expanded to incorporate heterogeneous age and cohort effects.

Each of the methods presented above can potentially be motivated by economic theory or by observing empirical realities by which to derive identification restrictions. Most often these assumptions go untested, however. McKenzie (2006) offers an approach that allows researchers to explicitly test whether sections of any of the APC profiles can be modelled with a linear or quadratic trend. If any portion of one of the APC profiles can be modelled parametrically, the rest of the model can be estimated by exhaustive dummy variables. While the trajectories and slopes of each profile cannot be uniquely determined, the acceleration of each is independent of restrictions imposed. Hence, shapes are unique, and structural breaks in the functional form are testable. An initial estimate of an additive APC equation can be obtained using any method above, and then Wald tests can be conducted to establish whether profile deceleration or acceleration changes for portions thereof. If the rate of acceleration changes, a structural break is discerned. If for any two consecutive ages, periods or birth years, the (unique) rate of acceleration does not change, a quadratic functional form would adequately define the section. This assertion can be tested for multiple sections along the profiles with a Wald test of the following null hypothesis:

$$H_0: quadratic\ profile \iff (\beta_{ij} - \beta_{ij-1}) - (\beta_{ij-1} - \beta_{ij-2}) = (\beta_{ij-1} - \beta_{ij-2}) - (\beta_{ij-2} - \beta_{ij-3})$$

$$\text{where } i = age, period, cohort;\ j = 1 \dots N_i \quad (4.4)$$

A follow-up test applies a similar procedure, but assesses whether the profiles can be restricted to linear portions. Linear profiles always exhibit zero acceleration, so that

$$H_0: linear\ profile \iff (\beta_{ij} - \beta_{ij-1}) - (\beta_{ij-1} - \beta_{ij-2}) = (\beta_{ij-1} - \beta_{ij-2}) - (\beta_{ij-2} - \beta_{ij-3}) = 0$$

$$\text{where } i = age, period, cohort;\ j = 1 \dots N_i \quad (4.5)$$

If both hypotheses are rejected for any section of the profiles, dummy variables are introduced for the non-linear and non-quadratic sections.

The benefit of this approach is that one needs to find only one such restriction, even if a number of them are statistically valid. If the statistical conclusions also concur with economic observations, then the combination of these criteria offers a solution that is satisfying along multiple dimensions. This approach presents an alternative that is well-motivated and allows the researcher to apply flexibility in their choice of structure, rather than atheoretical approaches (discussed below) or unverifiable behavioural assumptions[27]. Note, however, that implicit identifying assumptions do also apply: despite the unique determination of the acceleration of profiles, it is impossible to determine whether observations at the starting point of any profile have undergone acceleration, so that all tests are conducted on a restricted portion of the data.

Can this approach accommodate interactive profiles? A simple implementation could interact the sub-regions of the age and cohort profiles identified by this algorithm to incorporate heterogeneity. While it is conceptually possible to calculate directional second derivatives (in other words, to find the acceleration of profiles conditional on a value of another component, in order to allow for heterogeneity), the utility of this approach might be limited if similar structural breaks occur for all groups. For instance, a break in period effects is unlikely to differ by cohort or age group if it is a truly aggregate macroeconomic observance. Hence, testing for the various segments in an additive setting may be sufficient. The combination of testability of behavioural assumptions and the possibility of introducing heterogeneous profiles makes this an appealing approach.

4.2.2.2. Atheoretical approaches

Given the difficulty of contextually motivating some of the identifying restrictions proposed in the literature, purely statistically motivated approaches have been pursued as an alternative. One of the

[27] Because some discretion on the part of the researcher is required – such as identifying a break when the statistical evidence is marginally rejected, while economic evidence supports it – this approach is not investigated in the simulation study presented below.

most cited and widely applied estimators for an additive model is embodied in the so-called "intrinsic estimator" (IE) (Yang et al., 2004; Yang et al., 2008). It does not require the researcher to make behavioural assumptions about the data generating process, nor does it require parametric curve fitting for sub-portions of the profiles. However, the estimator imposes data-specific geometric conditions through the singular value decomposition, which are independent of context-specific theory.

Practical implementation follows a principal components analysis on the raw data (as opposed to a covariance matrix). However, by the singular value decomposition, the non-invertible design matrix of rank $k = A + P + C - 3 < A + P + C + 1$ can be re-written in terms of $\boldsymbol{u}_i$, the eigenvectors of $X'X$ (which together form the columns of a matrix U) and $\boldsymbol{v}_i$, the eigenvectors of $X'X$ (which together form the columns of a matrix V) and a matrix with only "diagonal elements" that are derived from the square root of the eigenvalues of $X'X$ (λ_i) (Johnson & Wichern, 2002, p.101). This matrix X can also include the base categories for each set of dummy variables in the model:

$$X = UDV' = \sum_{i=1}^{A+P+C+1} \lambda_i^{0.5}\, \boldsymbol{u}_i \boldsymbol{v}'_i$$

$$= \sum_{i=1}^{k} \lambda_i^{0.5}\, \boldsymbol{u}_i \boldsymbol{v}'_i + \sum_{i=k+1}^{A+P+C+1} 0 * \boldsymbol{u}_i \boldsymbol{v}'_i = \widetilde{U}\widetilde{D}\widetilde{V}' + \mathbf{0} \ldots \quad (4.6)$$

Because four of the eigenvalues are zero by definition, the final summation in equation 4.6 does not influence X, and therefore removes four eigenvectors (which are in the null space of X)[28]. As discussed in Appendix E, the new formulation $\widetilde{U}\widetilde{D}\widetilde{V}'$, which excludes each of the influences of all columns associated with the zero eigenvalues, can be used to construct the intrinsic estimator. In summary, one could therefore construct the design matrix using only the eigenvectors associated with non-zero eigenvalues. In effect, the matrix is now only dependent on a sum of $A + P + C - 3$ eigenvectors, which are also orthogonal to each other. Completing the APC analysis on only these eigenvectors (as opposed to the untransformed dummy variables) therefore eliminates all multicollinearity problems: the redundant vectors are excluded from the analysis, yielding a new design matrix of full rank. However, coefficient estimates have been rotated into another space (away from the null space of the design matrix) by the principal-component-like transformation. Hence, they should be rotated back into the standard APC space to enable usual interpretation. The rotation back into APC space implies that each of the effects is no longer orthogonal to each other, as was the case for the eigenvectors. As shown fully in Appendix E, the intrinsic estimator can therefore be

[28] Usually the literature refers to only one zero eigenvalue and one restriction on the design matrix. However, the approach of the intrinsic estimator also allows for the inclusion of the base categories of the dummy variables for age, period and cohort effects in estimation. Given that each set of APC dummies is collinear with the constant, these will also be associated with zero eigenvalues (the reason why they are omitted in usual OLS applications). This raises the number of eigenvalues to four if base categories are also included.

constructed using only the components of the singular value decomposition that do not correspond to the zero eigenvalues, as follows:

$$\widehat{\boldsymbol{\beta}}_{intrinsic} = \tilde{V}\tilde{D}^{-1}\tilde{U}'\boldsymbol{y} \ldots (4.7)$$

The intrinsic estimator will always differ from other estimators using behavioural restrictions, unless $P \to \infty$: the authors suggest an alternative geometric method to arrive at the solution to prove this (Yang et al., 2008). An initial step establishes APC estimates based on any of the coefficient restriction methods; then, that coefficient vector is projected into the non-null space of the design matrix, by removing the component associated with the zero eigenvalue (or the null space of the design matrix). The result of removing the influence of the null space from the estimator will always be the intrinsic estimator, regardless of the initial estimate. As a result, all additive APC estimates can be seen as a rotation of the intrinsic estimator, but with a different level of influence of the null space incorporated. Any APC estimator based on coefficient restrictions can therefore be expressed as:

$$\widehat{\boldsymbol{\beta}}_{apc} = \widehat{\boldsymbol{\beta}}_{intrinsic} + \alpha \boldsymbol{e}_0 \ldots (4.8)$$

where $\boldsymbol{e}_0$ is the eigenvector associated with the zero eigenvalue of the design matrix. Each behavioural restriction is represented by a different value of α, and determines the degree of rotation away from the intrinsic estimator. Contrary to most interpretations, the IE *does* impose restrictions in estimation. That is, the IE sets $\alpha = 0$, which does impose a problem-specific behavioural assumption that remains unknown to the applied researcher. Hence, it is a purely mechanical approach whose assumptions cannot be easily translated back into the real world with any ease. Because all estimators based on one parameter constraint are simply rotated variations of the intrinsic estimator, it is true that they will all yield the same fit of the data, and that their shapes will be identical (though slopes will not be), as confirmed by McKenzie (2006).

Yang et al. (2008) show that $\boldsymbol{e}_0$ depends uniquely and solely on the number of time periods and cohorts included in the data. By projecting estimates away from this space, this estimator therefore also supposedly "corrects" for the limited span of time periods, cohorts and age groups in the data: this information fully determines the null space of the design matrix, from which the estimator is projected away. Hence, using few surveys with the intrinsic estimator is supposedly not as serious as when the required assumption is that period effects should represent business cycle variation (Deaton, 1997). In fact, Yang et al. (2008) maintain that as the number of time periods increases, the estimates obtained by way of alternative constraints converge to the intrinsic estimator, so that $\alpha \to 0$, as $P \to \infty$, regardless of identification strategy. By this asymptotic argument, the IE is supposedly consistent, though subsequent reviews dispute this (Luo, 2013).

Because the IE restricts the influence of the null space of the design matrix, and not coefficients explicitly, it is difficult to compare its results with those of other methods. Yang et al. (2008) suggest that many authors are not aware of the implications of using the IE, because it is not associated with easily understandable behavioural assumptions. While, for instance, Deaton's (1997) restriction is transparent about the fact that period coefficients only yield transitory movements, no analogous interpretation of the restrictions of the IE can be provided. Many authors assume that using the IE is less restrictive compared to traditional methods, but this is not true. In reality a restriction is still imposed, though it is now more abstract in nature. The researcher has not imposed the identification assumption, whilst the principal components' transformation has done so on his or her behalf, without providing clues as to what the behavioural assumption it is that has been adhered to.

Instead of providing unbiased and consistent estimates, the intrinsic estimator has been classed alongside other restrictions in their inability to additively separate APC effects (Luo, 2013). Simulation evidence that sets up a data generating processes that closely matches $\alpha = 0$ does confirm that the intrinsic estimator is the better option compared to traditional methods (Yang et al., 2008). However, simulations that implement data generating processes that more closely approximate other behavioural assumptions show that the intrinsic estimator is inferior in obtaining true effects (Luo, 2013). Given that all of these methods can be placed in the same class - in that each allows for a different degree of influence of the null space – estimates are only valid if the correct identifying assumption has been chosen. However, it is not usually clear which identifying assumptions are valid in each scenario. Hence, the intrinsic estimator is not uniquely superior nor unequivocally bias-free, and is simply a rotation of results obtained from other approaches.

The fact that the intrinsic estimator does not remove all bias in all circumstances is explicitly acknowledged by the original authors, and they suggest that it be applied (as with any identification problem in econometrics) with due consideration of the population process that is likely to underlie the data (Yang & Land, 2013). However, given that it is difficult to connect behavioural assumptions with the underlying mechanics of the intrinsic estimator, it is not clear that this approach can be adequately motivated in many circumstances.

The intrinsic estimator has recently been applied to study the evolution of returns to education in South Africa, motivated by the correct assertion that alternative behavioural assumptions cannot be justified (Branson et al., 2013). For instance, returns to education are unlikely to be cyclically determined, so that the Deaton (1997) restriction is not appropriate. However, it is also not certain that the implicit behavioural assumption made by the IE estimator matches reality. Hence, this "assumption agnostic" approach can in effect not be classed any differently to others, and does not present a more workable solution to the APC identification problem (Fienberg, 2013).

Despite not being a superior estimator, the IE is however an ideal candidate to model APC effects interactively. No known application extends the intrinsic estimator beyond the additive model. In fact, most existing user-written software prevents researchers from fitting heterogeneous APC models. However, it is mathematically feasible to extend the arguments for additive modelling to interactive effects. While interactive effects will yield another host of zero eigenvalues in the design matrix, their associated eigenvectors can be excluded from estimation in the same manner as in the additive case.

Browning et al. (2012) also move to statistical criteria rather than behavioural restrictions to identify the APC components. However, they are explicit about the fact that APC effects cannot be point identified, but are more suited to set identification. They set out to develop a method based on the maximum entropy principle. This approach searches for a probability distribution of the APC estimate that has maximum entropy (in other words using at most the information available in prior data, assuming that the rest is random), and is a useful alternative when a problem is underidentified. The notion of maximum entropy is based on LaPlace's "principle of insufficient reason", suggesting that one should choose the distribution for estimates that is most conservative, given the limited available information for identification (Mittelhammer et al., 2000, pp.E3:10). For instance, if one is to predict the next roll of a die without any prior data, the most conservative distribution for an estimate of the next outcome would be uniform with a probability of 1/6 for each outcome. However, with prior information, it is possible to choose a distribution that places more weight on some outcomes than others. It has more structure, or less entropy, but not in a manner that the data is unable to identify. Entropy is maximised, so that no more information is inferred than the data allows.

The approach therefore maximizes a measure of entropy, subject to the constraints of the problem. A common measure of entropy used by many authors was originally proposed by Shannon (1948) as $-\sum_{i=1}^{k} p_i ln(p_i)$, where $0ln(0)$ is assumed to equal 0, and where p is the probability that each estimate in a vector space is the true population value. Following Browning et al. (2012) the APC estimation can be reparametrised as $\widehat{\boldsymbol{\beta}}_{apc} = A\boldsymbol{\beta} = AS\boldsymbol{p}$, where A represents a matrix that imposes restrictions on the population parameter vector to yield the reduced form estimate $\widehat{\boldsymbol{\beta}}_{apc}$, and S is a set of vectors that defines the space of all possible estimates of $\boldsymbol{\beta}$, the population parameter, with $\boldsymbol{p}$ representing a vector of probabilities that each of the vectors in S is the real solution. The process of finding $\widehat{\boldsymbol{\beta}}_{apc}$now amounts to solving a Lagrangian that maximizes entropy, subject to the constraint that $\widehat{\boldsymbol{\beta}}_{apc} = AS\boldsymbol{p}$ and that probabilities must be positive and sum to 1. Without constraints, the solution would assign equal probability to all possible estimates. However, the APC problem constrains this solution, with the result representing the most conservatively informative distribution of estimates. The expected value of this distribution then represents the maximum entropy APC estimate. The authors conclude (on the basis of empirical examples) that the strategy successfully recovers APC coefficients, though they make no claims that this is a general solution to the identification problem, as it simply weights

all possible solutions (Browning et al., 2012). While this approach can be potentially extended to accommodate heterogeneous age profiles, this option is not further investigated in this context, given further weaknesses identified with this method in the simulation study below.

Given the large collective evidence that points towards the impossibility of parametric identification of the APC problem, the most recent attempts have shifted the attention to non-parametric techniques (Jiang & Carriere, 2014). The basis for this approach is the so-called generalized additive model (GAM), which was developed by Hastie & Tibshirani (1990). The next section will expand on GAMs as a potential solution to interactive APC problems, but the basic notion behind estimation of the additive model is introduced here. GAMs are more flexible than those that impose behavioural assumptions, in that they model a set of smoothed curves of the outcome variable along the distribution of predictors. In the case of APC models, a non-parametric form could be conceptualized as:

$$y_{apc} = \mu + f_1(a) + f_2(p) + f_3(c) + \varepsilon_{apc} \dots (4.9)$$

It is simple to estimate each of the f_i separately using bivariate local polynomial or lowess smoothers (for instance), but this would defeat the object of holding all other factors constant, as is the case in linear models. While each of the functions should represent partial relationships between the APC components and the outcome of interest, this objective is not simply implemented by a single non-parametric estimator as is the case in parametric specifications. Hence, iteration is required, as embodied by the so-called backfitting algorithm. An initial estimate of one of the bivariate estimators of f_i is implemented. Residuals from this result are then used as the input for another bivariate estimate of one of the other variables. The residuals are then repeatedly modelled until the estimates converge to stable values. While this procedure seems arbitrary in its sequence, Hastie & Tibshirani (1990) show that estimates are stable and have the ability to generate expected results, with unique identification in some circumstances. In the presence of *concurvity* (the non-parametric analogue to perfect multicollinearity), however, the starting estimator matters for each of the resulting curves individually, but not for the overall prediction (Hastie & Tibshirani, 1990, p.120). This means that APC profiles estimated by GAMs are also not unique. In fact, the authors offer a solution that is very similar in spirit to the intrinsic estimator, in that they proceed with the backfitting algorithm on variables that are projected away from the null space. GAMs may be the most flexible manner to attempt to identify the APC effects, but also do not present a general solution to the problem.

However, one distinct advantage of GAMs is that the model can be adapted to include interactive effects, even if these are not unique:

$$y_{apc} = \mu + f_1(a; c) + f_2(p) + \varepsilon_{apc} \dots (4.10)$$

The approach is not very different from the additive counterpart, except that f_1 estimates a three-dimensional surface in each of the backfitting iterations by (for instance) a lowess smoother. Importantly, predictions from this model generate a different age profile for each cohort; and, by implication, it also ensures a variable generational trajectory at each age. Generalising this to triple interactions is not possible: in any event, it would amount to a perfect fit, with each *apc* cell being predicted by its own average value. Similarly to the additive model, which requires a restriction on one of the variables, it is necessary to limit one set of interactions. In the application below no interactions with period will be included, though they too could be important elements of the data generating process.

Based on simulation evidence, Jiang & Carriere (2014) conclude that the additive GAM approach delivers lower root mean square errors than behavioural restrictions for linear APC estimates, and is therefore likely to reflect the population estimate more often. These findings are especially true where the cells of an APC analysis are aggregated from micro data, and some cell outcome statistics are constructed with few underlying observations. Such concerns are relevant when endpoint cohorts do not appear in the data as readily (such as those who are born in early years, who have already exited the labour market in large numbers).

Keyes & Li (2010) propose that interactive effects can be included implicitly in additive models. They suggest conducting a first stage estimate by Polish median estimation (Tukey, 1977), whereby only a two-way decomposition is conducted:

$$y_{ap} = \alpha_a + \pi_p + \varepsilon_{ap} \ \ldots (4.11)$$

Since cohort effects are omitted from this analysis, they are implicitly included in the error term as an interaction of age and period effects. A dummy variable regression on the Polish median residuals therefore generates a cohort profile without trouble. Because Polish median estimation is not based on a linear model, it is possible to use a standard OLS regression in the second stage. The benefit of this two-stage procedure is that it can easily be adapted to account for variable age profiles (for a subset of the full number of periods) without a collinearity problem in the first step, as follows:

$$y_{ap} = \alpha_a + \pi_p + \delta_{ap} + \varepsilon_{ap} \ \ldots (4.12)$$

However, a distinct limitation is the sequence of estimation. The first step does not condition on the omitted factor (cohort), so that those profiles are unlikely to reflect true age profiles net of generational effects. It is furthermore possible to change the estimation sequence substantially, by interchanging the variable couple included in the first stage of the model. The implied path dependence could therefore be influential for the final conclusions reached. In essence, this approach is equivalent to an over-simplified GAM (Hastie & Tibshirani, 1990) that stops at the first backfitting

iteration, without taking into account the full set of controls in the entire procedure. Hence, an implementation of a GAM is the minimum that should be pursued in such a setting; however, as discussed below, it can also be implemented to find heterogeneous APC effects. In this study, a semi-parametric two phase approach is also followed, based on the reasoning of Keyes & Li (2010). The first stage is run by an OLS regression conditioning only on age and period dummies, while a local polynomial smoother fits the cohort profile from the residuals. However, it is not expected that this approach will yield many gains relative to a GAM, which takes full cognizance of partial correlations. Hence, the interactive solution is omitted.

Another path dependent approach views age as the unique outcome of an individual's year of birth and the current year, lending itself to hierarchical linear analysis, with age effects modelled at a "macro" level, and other effects at a micro level[29] (Yang & Land, 2006). However, simulation evidence also concludes that this method is not robust to all types of data generating processes, and hence also does not present a universal solution to the APC identification problem (Bell & Jones, 2014). This method will not be discussed further or investigated in this study.

4.3. Monte-Carlo simulation evidence

4.3.1. Data generating processes

This section investigates the plausibility of implementing various APC estimators, in settings where the data generating process is firstly additive, and secondly interactive. The first scenario assumes that cohort effects are permanent over the life cycle, while the second relaxes this assumption. In particular, the second scenario is informed by the evolution of South African labour force participation, which has been subjected to extensive cohort analysis (Branson & Wittenberg, 2007; Burger & von Fintel, 2014; Burger et al., 2014). One of the salient features of labour force participation has been the progressive integration of blacks into the job market. This can be characterized by a long-run permanent cohort trend, driven by the pull of progressively increasing wages for black workers and the abandonment of apartheid segregation policies (Burger & von Fintel, 2014). Alternatively, this can be characterized as a changing age profile, with more recent generations entering the labour market at earlier ages (Branson & Wittenberg, 2007) likely due to new education policies that limited the number of years learners are allowed to stay in school (Burger et al., 2014). The simulated process mimics the propensity of an individual to participate in the labour market, though it is more generally representative of a problem where cohort effects are non-permanent. The goal of the Monte-Carlo simulation study is to understand the implications of using additive models

[29]For instance, in school production functions that use hierarchical models, the macro unit of observation is represented by a school, while the micro unit is the learner. Here age is seen as a "higher" unit of observation, with period and cohort each representing sub-units that contribute to this level, because age is uniquely determined by an individuals' generation and the period in which the data is recorded.

when the underlying process is truly interactive, and to also identify suitable methods for estimating heterogeneous APC profiles.

An interactive data generating process is simulated based on indicative values from Labour Force Survey data. The additive version simulates data that restricts D to be zero for all observations:

$$p^* = -0.75 + 0.075age - 0.0009375age^2$$
$$+D(\,0.0005age - 0.00001725age^2 + 0.005\,)(birthyear - 1966)$$
$$+0.0075birthyear - 0.05\cos\left[2\pi\left(\frac{year-1995}{2010-1995}\right)\right] + e \quad ...(4.13)$$

where $D = I(birthyear \geq 1965)$ and $e \sim N(0; 0.01)$; *year*=1995…2010; *age*=15…65; *birthyear* = 1930…1995

The first group of terms in equation (4.13) represents a standard concave age profile. However, in the second set of simulations (where $D \neq 0$), this is specified interactively (as is evident in the second group of terms), with birth cohorts born after the middle of the cohort range experiencing a shifting age profile. Implicit in this shift is a gradual positive linear cohort trend of 0.005 that is specific to these cohorts (over and above other cohort trends). Furthermore, an adjustment in the shape of the age profile is imposed for more recent cohorts: the positive interaction on the age coefficient entails higher values of the dependent variable at earlier ages for newer generations, though the dampening effect of the quadratic is also larger. The third group of terms represents a linear cohort trend that affects *all* generations, with a premium of 0.0075 for each year of birth, as well as a cyclical period effect modelled by a *cos* wave.

Figure 4.2 includes a visual depiction of the additive simulated data, while Figure 4.1 shows the interactive simulated data. The additive data is simple to interpret, with a concave age profile that is not differentiated by generation, a permanent cohort effect that rises linearly over birth year, and a period effect representing one full business cycle. The interactive data requires more in depth explanation. The bottom left panel of Figure 4.1 shows age effects, with each line representing the life cycle effect of a specific birth cohort. The left-most visible line represents those born in 1990. Each line to the right represents a group born 5 years later, proceeding all the way to the 1930 cohort. The interactive term in the data generating process ensures that younger generations have a higher participation profile than slightly older generations, but only early in the life cycle. This is achieved by the positive interaction effect on age, while the negative on the square term ensures a dampening effect, so that later in the life cycle all birth cohorts follow the same pattern. Such a process could be interpreted as (and estimated as) cohort effects, but they cannot be interpreted as permanent across the life cycle, with differences only existing in early phases. This setup purposefully defies the

assumptions of the additive model, which implicitly requires that "formative experiences" remain permanent across the life cycle (Glenn, 1976).

The bottom right panel of Figure 4.1 depicts cohort profiles. The "average" trajectory that stretches across the entire birth year domain shows a linear cohort trend that is common to all generations up until 1965 (which is represented by the third set of terms in equation 4.13). All later birth years have a steeper profile, which is effected by the constant of 0.005 in the interaction term. However, cohort profiles can also be viewed as age-specific, due to the interaction with this variable. Each of the other lines shows cohort profiles for different ages: they essentially recast the data presented in the age profiles, but condition each line on age. They, however, incorporate both the "average" component discussed before *and* the age effect. While it seems superfluous to present the profiles that confound the age and cohort effects in this manner, it becomes clearer in the estimation process that this is a useful step. The reason is that one never obtains "one" average cohort profile from interactive estimates, as they depend on age. While it is possible to detrend these specific cohort profiles of average age effects, cohort profiles nevertheless remain heterogeneous by age group. It is therefore necessary to average the age-detrended effect over cohorts to approximate a "single" cohort effect.

Finally, the last group of terms in equation (4.13) adds a cyclical time effect with an upward phase followed by a downward phase. This is depicted in the top right panel of Figure 4.1.

The top left panel of Figure 4.1 summarizes the entire data generating process in a contour plot. Moving along the age axis, it is evident that more recent generations have higher values of the outcome variable earlier in life. The score is highest at prime ages, as was also evident in the age profiles. The distinct "diagonal" pattern represents period effects peak in about 2002. Furthermore, the overall pattern is not exactly symmetric, because younger generations enter the labour market earlier.

Microdata is simulated independently for 16 periods (in order to roughly match the survey years for the empirical application) with 25000 observations per wave. In each period, an age distribution is generated according to a lognormal distribution (assuming a relatively youthful population), where $\log(age) \sim N(3.425; 0.16)$. This distribution roughly agrees with Labour Force Survey data. Birth years are then implicitly calculated by the APC identity. The generated outcome values are then censored to a binary variable, where all observations above the mean in a particular year are assigned a value of 1, and 0 otherwise. Data is then collapsed into *apc* cells, representing probabilities of a

positive outcome for each age-year-birth year combination[30]. In addition, each simulation adds an error term randomly drawn from a normal distribution of mean zero and standard deviation 0.1.

While it is evident *a priori* that additive models cannot solve for the interactive data generating structure, simulations test whether any of the standard approaches can approximate the desired solution on both the additive and interactive data. The comparison between the two procedures clearly illustrates the potential dangers of assuming an additive structure when the population does in fact contain heterogeneous APC effects. As the simulations illustrate, temporary changes to the age profile are generally estimated as permanent cohort effects in additive APC models, so that additive models can be potentially misleading. Simulations include applications of the intrinsic estimator (Yang et al., 2008), the Deaton-Paxon restriction (Deaton & Paxon, 1994), the maximum entropy estimator (Browning et al., 2012), an adaptation of the multi-phase semi-parametric model of Keyes & Li (2010) and also a GAM without interactions (Jiang & Carriere, 2014). Finally, the GAM and IE are adapted to allow for fully interactive age and cohort profiles on the interactive data, in order to illustrate potential gains from flexible models that are not bound by the restrictions on the design matrix. While the structural break tests of McKenzie (2006) present useful alternatives for interactive specifications, they are only implemented on the real data, as some discretion based on economic knowledge is required in their implementation.

The Monte-Carlo simulation study proceeds by taking 1000 draws from the proposed data generating processes. Relevant estimates are calculated using the data generated during *each* draw. The discussion below focuses on the average of estimates from each of the 1000 calculations. Appendix F presents the distribution of estimates from each Monte-Carlo simulation.

[30] Results using individual data for the binary outcome variable are very similar, but computationally more intensive, especially where non-parametric techniques are utilised. Hence, the smaller dataset which takes the *apc* cells as the unit of observation is utilized, without loss of generalization.

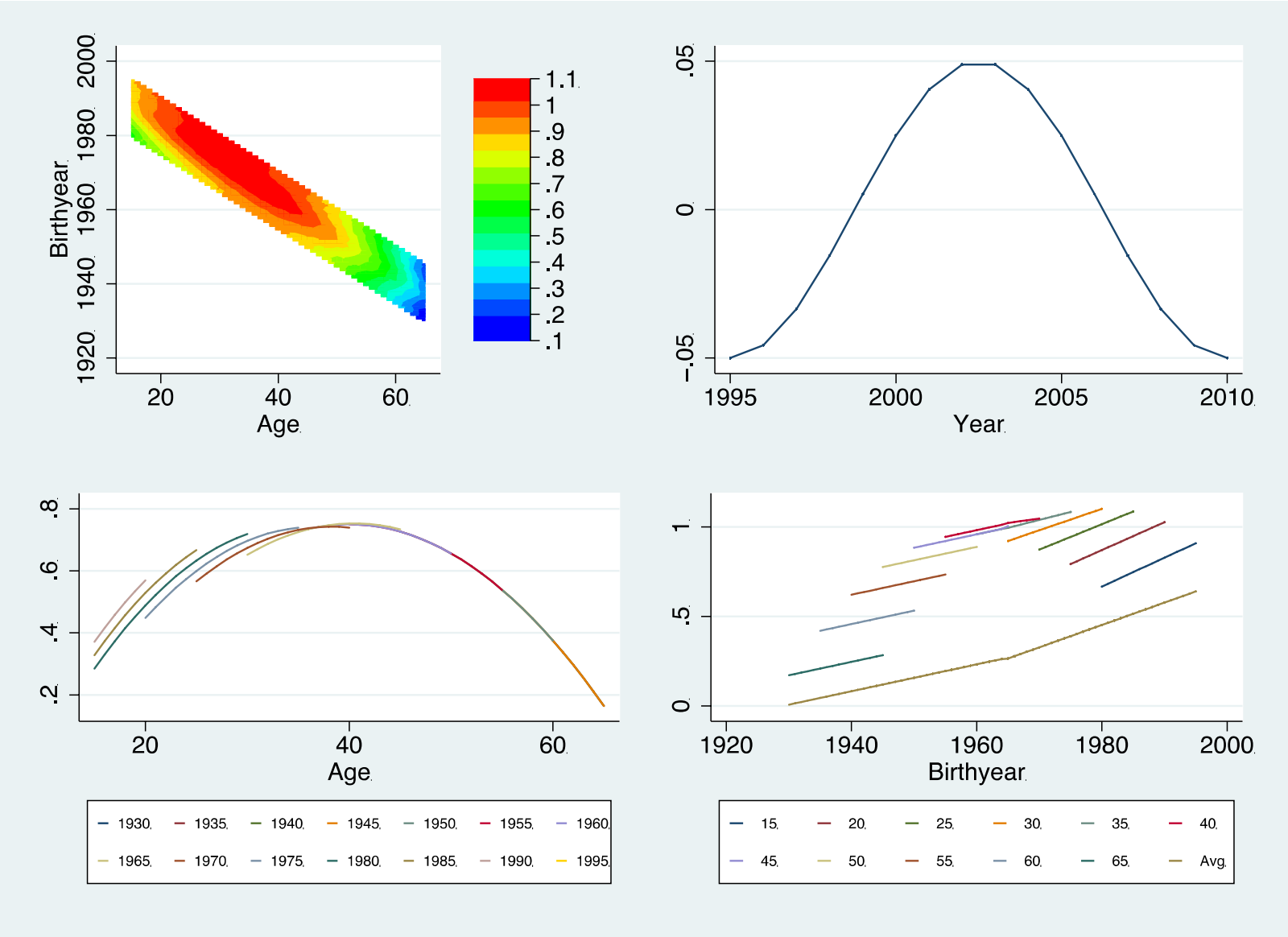

Figure 4.1 Depiction of the simulated interactive data generating process (Equation 4.13)

4.3.2. Age-period-cohort models applied to simulated data

4.3.2.1. Additive models on additive data

Before discussing the implications of applying additive models to data that is actually interactive, the APC estimators are calculated on the simulated data generating process that is restricted to be additive ($D = 0$ in equation 4.13). The top right panel of Figure 4.2 summarizes age profiles. While each is concave, no estimator gets the shape right. In particular, they are flatter at the beginning and end of the life cycle than the simulated data would suggest, and more peaked in prime ages. Those who were aged 58-65 in 1995 experienced a decrease in the outcome variable due to the life cycle effect in the simulated data, but were also subject to an increase due to the business cycle effect. All of the estimators confuse these effects, as they cancel out in the age profile. Additionally, most estimators capture a zero period and cohort trajectory, supporting that these effects are confounded and captured in the age profile instead.

The intrinsic and maximum entropy estimators are the only ones to yield cohort and period profiles that are not flat. The IE is the only estimator to trace out a positive (yet largely muted) cohort trajectory as contained in the simulated data, but it estimates only a linear downward period effect. The maximum entropy estimator does the opposite, emphasizing a downward cohort profile and an upward period effect. Here it is evident that there is a trade-off in the estimation of period and cohort effects, with no estimator able to recover the cyclical time variation from the long-run (permanent) cohort changes. Simulated results therefore show that even when the data generating process is really additive, APC estimators are poor at unconfounding each of the effects, especially when time variation is truly cyclical and all long-run change is in reality represented by a cohort effect.

Why do these estimators perform so differently? Given that only the Deaton-Paxon estimator implements explicit behavioural assumptions, it is not immediately clear why this is the case. This estimator has averaged the period effects to zero, but in doing so has enforced *each* period to assume this value. This is not always the case, as the estimates of Burger & von Fintel (2014) show. The semi-parametric estimator is weak, in that the first stage does not condition on cohort, but rather confounds all effects in the age profiles. The GAM improves by providing conditional profiles through the backfitting algorithm, but this appears to be inconsequential for conclusions. The intrinsic estimator and maximum entropy estimator apply purely statistical criteria, and do perform better at unconfounding the effects, even if

imperfectly. In the case of the former this is possible because the null space is excluded from the estimator, removing the source of multicollinearity. In the case of the latter, the procedure searches over all possible estimates, downweighting those that confound the profiles. The reason why these two estimators deal with the trade-off between period and cohort effects so differently is not clear. Nor is the explicit relationship between these estimators and the Deaton-Paxon estimator self-evident, beyond noting that all three estimates have the same shape and that they are rotated versions of each other (as in equation 4.7.). Without further knowledge about how the null space influences behavioural assumptions, the relationship between estimates is also not pinpointed.

4.3.2.2. Additive models on interactive data

This section illustrates how temporary cohort effects are absorbed into APC profiles by additive models. Figure 4.3 summarizes the *additive* APC results from simulations of the *interactive* data generating process[31]. The top left panel repeats the interactive age-cohort structure that was simulated. The top right panel shows the average age profile obtained by the various additive estimators. Again, each estimator generates an approximately concave age profile. However, not one case can account for the highly flexible interactive age structure that is central to the simulated data, yielding similar results compared to the additive simulated data. Additionally, Figures F.1 to F.5 in Appendix F also show that for many estimators, the distribution of simulation-based estimates is narrow for the early part of the life cycle compared to the last part. This contrasts with the fact that the simulated data actually "fans out" early in the life cycle, where the cohort heterogeneity is introduced into the age profile. Hence, the interactive effects must have been captured by other profiles.

Cohort profiles differ substantially from those using only the additive data, capturing all the heterogeneity that interaction introduced. As in the additive case, profiles remain muted before the changing slope introduced for cohorts born after 1965. Thereafter, the structural break that is built into the data generating process at this birth year is picked up by all estimators, bar for the GAM and semi-parametric approaches. Hence, as noted before, age and cohort effects are clearly confounded in most estimators. The best performer on the cohort profile is again the intrinsic estimator, which displays a small upward portion initially, but the downward section in the centre of the cohort profile means that this estimator is not entirely successful. By implication, if additive estimators are applied to data that is really interactive, cohort effects that are only relevant for a part of the life cycle are reflected as permanent. This has implications for existing evidence on generational labour force

[31] The appendix presents a full set of results, where box plots give an indication of the spread of the estimators under question.

participation in South Africa (Burger & von Fintel, 2014), and will therefore be re-investigated below.

As in the additive case, all estimators fail to capture the cyclical component inherent to the data generating process. The maximum entropy estimator only captures the upward phase of the period effects, while it yields a strong downward cohort profile. The intrinsic estimator does the opposite, emphasizing the downward section of the period profile. Similarly, the semi-parametric analysis uncovers a pattern that resembles the intrinsic estimator, though the effects are far more modest. The Deaton-Paxon restriction again ignores the period effects altogether, whilst estimating a strong downward cohort effect for those born before 1965, but which climbs after that point.

Finally, while the work of Jiang & Carriere (2014) suggests that GAM models are consistently estimated even in the presence of sparse data, it is however true that the procedure is sensitive at endpoints of profiles (see Figure F.5 in Appendix F). Cohort and period profiles have consistently broader ranges of estimates at their endpoints for the additive GAM model. Endpoints can usually be associated with sparse regions in the data, where (for instance) cohorts that exit or enter the labour market are not found in as many surveys as other groups. However, this does not appear to be the problem here, because period effects also exhibit a wide range, and the simulated data contained the same number of underlying observations at each point in time. Rather, the non-parametric implementation is sensitive at the *edges* of the data, where smoothing is sensitive to "corner observations". This flexible approach may therefore be unsuitable to use in APC analysis, contrary to the conclusions of Jiang & Carriere (2014).

In summary, it is evident that incorrectly applying additive models to interactive data yields cohort profiles that are more pronounced, and incorrectly interpreted as *permanent* generational change. In all cases at least one section of one of the profiles was confounded with one of the other effects. However, some estimators perform better than others. The intrinsic estimator captures a basic age profile together with the cohort trend (except in the middle ranges) and one half of the cyclical effect. While it does not (and cannot) capture all of the interactive features of the data generating process, it is the best performer. Importantly, however, the Deaton-Paxon and maximum entropy profiles can be observed to be simple rotations of this estimator. This confirms the assertion in equation (4.7) that restrictions represent the intrinsic estimator plus the variable influence of the null space of the design matrix. It furthermore suggests (though this is not verified formally), that the mechanics of the maximum entropy estimator also operate as a restriction on the design matrix, despite being a probabilistic combination of all possible estimators. Similarly to the intrinsic

estimator, the maximum entropy principle recovers APC effects that rely solely on statistical principles, but which are associated with real but unknown behavioural restrictions on the model. While each of these estimators present appealing alternatives to identifying APC effects, it is not certain how to interpret the *behavioural* restrictions imposed by each.

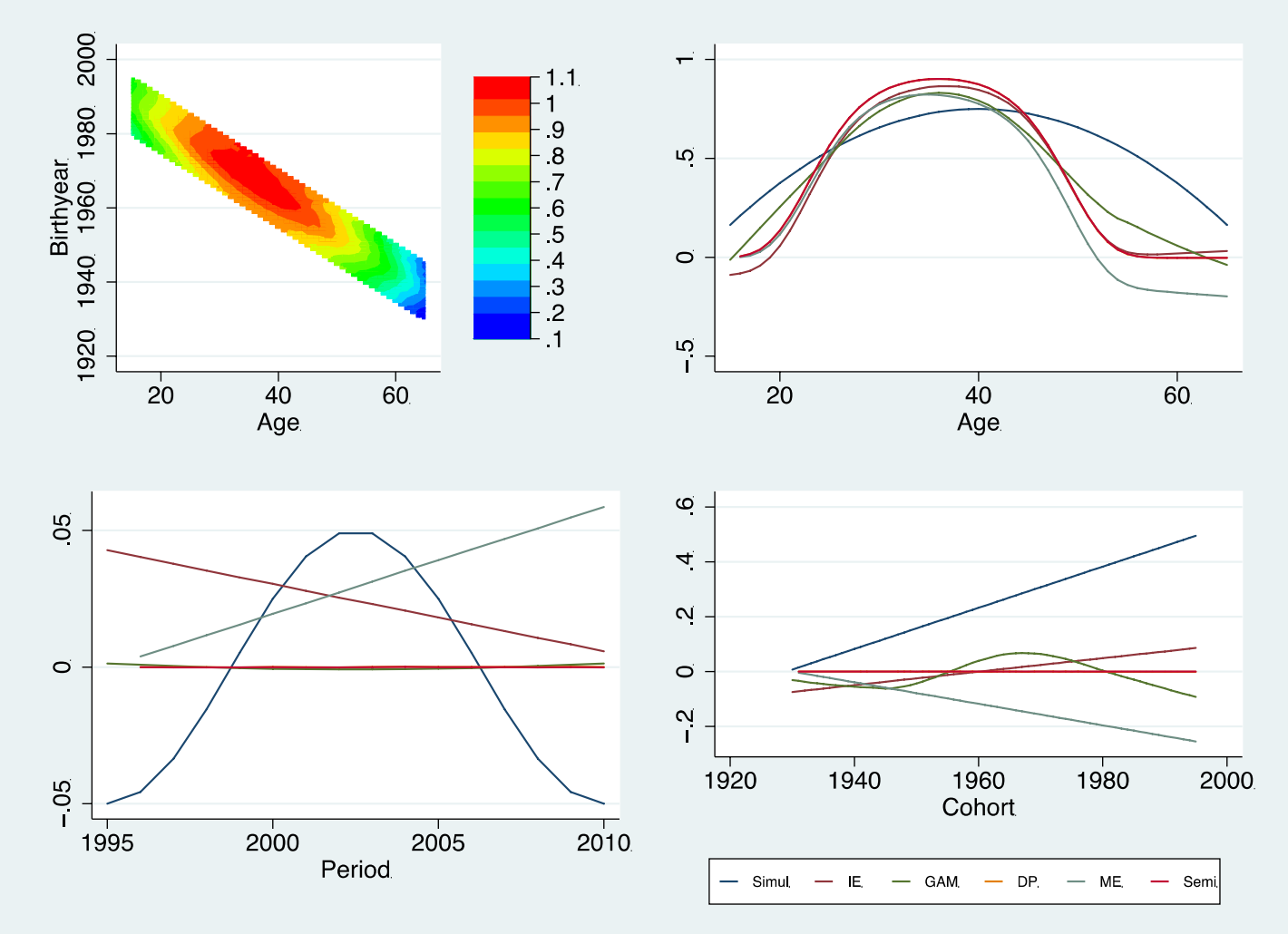

Figure 4.2 Simulated additive data generating process with mean additive APC estimates.

NOTES: Own calculations from 1000 Monte-Carlo simulations of equation 4.13 with $D = 0$. "Simul" = Simulated Data, "IE" = Additive Intrinsic Estimator, "GAM" = Additive Generalized Additive Model; "DP" = Deaton-Paxon restriction; "ME" = Maximum Entropy estimates; "Semi"= Semi-parametric estimator

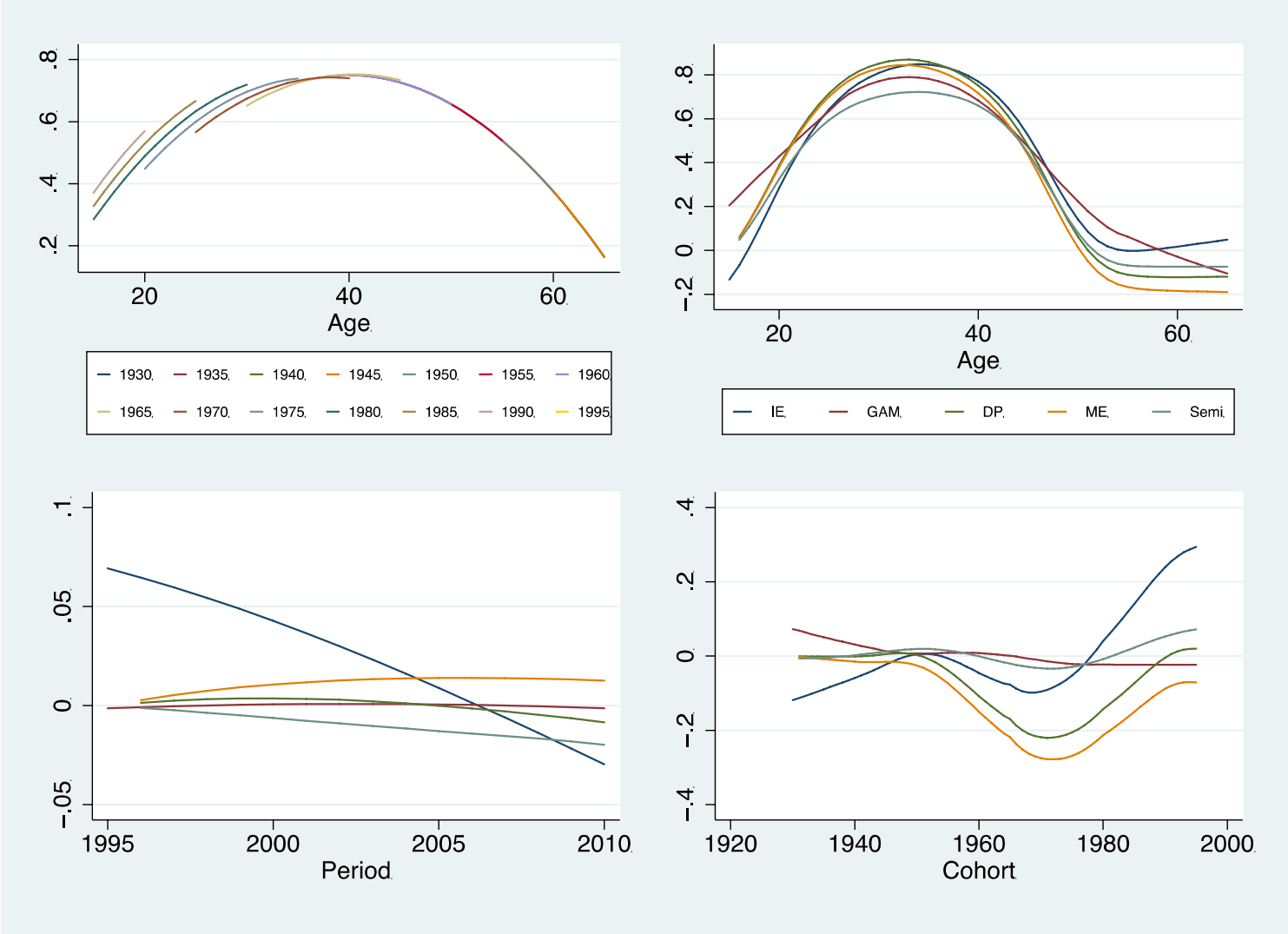

Figure 4.3 Mean APC profiles of additive models using interactive simulated data

NOTES: Own calculations from 1000 Monte-Carlo simulations of equation 4.13 with $D \neq 0$. "IE" = Additive Intrinsic Estimator, "GAM" = Additive Generalized Additive Model; "DP" = Deaton-Paxon restriction; "ME" = Maximum Entropy estimates; "Semi"= Semi-parametric estimator

4.3.2.3. Interactive models on interactive data

It is therefore necessary to investigate the possibility of non-permanent cohort effects that are embodied by interactive age profiles. Figure 4.4 compares the simulated interactive data generating process to mean estimates obtained from interactive Generalized Additive (GAM) and Intrinsic Estimator (IE) models. The former estimates age and cohort effects as a three-dimensional surface, and period effects as a bivariate curve, while the latter applies the principal components regression to a design matrix that includes exhaustive interactions between age and cohort dummies. Based on the evidence in the previous section, a priori expectations suggest that the IE variant of the interactive model could provide better results.

The first row of graphs represents the simulated data, while the second row shows the results obtained from the application of the interactive GAM, and the third row reports on the interactive IE estimator. Starting with the age profiles in the third column, the utility of the interactive model becomes immediately clear. The estimates pick up the higher profiles for recent generations at young ages. While the magnitudes from IE estimates start at values that are lower than the simulated data at the beginning of the life cycle, and are too high by age 40, it appears to pick up the "fanning" in the age effects at an appropriate point on the profile. While the life cycle progression at the beginning of the profile is better captured by the GAM, the effects of fanning are somewhat muted relative to the IE at this point, and therefore capture the heterogeneity with less success.

Some other important deviations from the simulated data also occur. Firstly, the peak of the age distribution is subject to substantial variation across cohorts (with a "thick" band of estimates in both cases). However, the effect should have been identical across cohorts immediately after the turning point at age 40, as in the simulated data. While the estimate does pick up the "fanning" (representing higher values for more recent birth years) implicit in the data generating process at young ages, this also occurs at and after the turning point of the age profile, rather than only before it. Some of these anomalies become clear in the discussion of period and cohort effects, which tend to confound the age profiles. Nevertheless, this representation of the model is the closest to recovering the true data generating process, and explicitly acknowledges the temporary nature of cohort effects.

The unexpected fanning at the turning point in the age profiles can potentially be explained by age's confoundedness with period and cohort effects. It is notable in the fourth column that estimated year effects from the GAM model have a shape that is very close to that of the simulated data. However, they are one tenth of the size compared to the simulated data. Given that the heat map in the top left corner highlights the concentration of period effects along the

diagonal (especially in the centre of the age distribution), the delayed "fanning" that is estimated in the life cycle is actually capturing a period effect (the peak of the cycle in the simulated data). Hence, the interactive GAM has not separately identified the confounding factors. Similarly to the additive IE model, the interactive IE has estimated only one half of the simulated cycle, but with a magnitude that is closer to the simulated effect than was the case for the interactive GAM. The large year effects, albeit without the cyclical pattern, entail that the estimator does not tend to mute the fanning effect just before the turning point. Hence, the IE is better than the GAM at separating age from year effects, but is poor at capturing cyclical variation.

Turning back to the first column, the simulated heat map is repeated in the first row, while in the second and third rows each line represents a cohort effect that is conditioned on age. Here age effects are not yet netted out, as the GAM model estimates the age-cohort surface in one step. The second column, however, differences away average age effects for each birth year to isolate pure cohort trajectories (that are nevertheless still different for each age group). The dashed black lines represent the average of these heterogeneous profiles, in order to recover the "permanent" cohort effect. The simulated data in the second column shows that cohort effects should differ conditional on age due to the interactive data generating process, but that "on average" a linear trend should be observed throughout, with a breakpoint for cohorts born after 1965. Turning again to the first column, both the GAM and IE estimators yield heterogeneous cohort profiles that are positively sloped for younger generations, though the slopes are steeper than the simulated data would suggest (particularly for the IE estimator). Once age effects are removed in the second column, the centre of the birth year distribution is characterized by U-shaped cohort estimates (rather than a linearly rising component), which represents a trade-off with the over-emphasized estimated age distribution in this region of the data, as discussed before. Furthermore, cohort profiles should rise linearly prior to the 1965 year of birth. However, they are downward sloping over this region using the GAM estimator, and only upward sloping for the very oldest generations using the IE estimator. The average trends also reflect these patterns. While neither the interactive GAM nor the interactive IE estimators have fully recovered the true cohort trend, the IE estimator has done so better. It captures an upward slope for the oldest generations, as well as the structural break in the trajectory. It is, however, the negative slope of each profile in the middle range that is incorrectly estimated, and can be explained by the persistent confoundedness with age profiles.

In sum, while perfect concurvity prevents the unique separation of heterogeneous APC effects, this section shows that it is possible to distinguish a permanent cohort trajectory from cohort trends that only affect the life cycle in sub-regions of the data. This innovation has

rarely been pursued in the literature. Despite the shortcomings noted amongst authors on additive IE models, in addition to their inability to distinguish short-run cycles, the IE is the most suitable candidate to pursue this route. One objection to this type of modelling might be its lack of generality, with many profiles being estimated. However, the combination of such estimation procedures with formal testing to delineate structural breaks (McKenzie, 2006), can deliver a method that provides heterogeneous profiles across broader regions of the data. This option is pursued on real data. Hence, this section has laid the foundation for appropriate empirical analysis of interactive APC models.

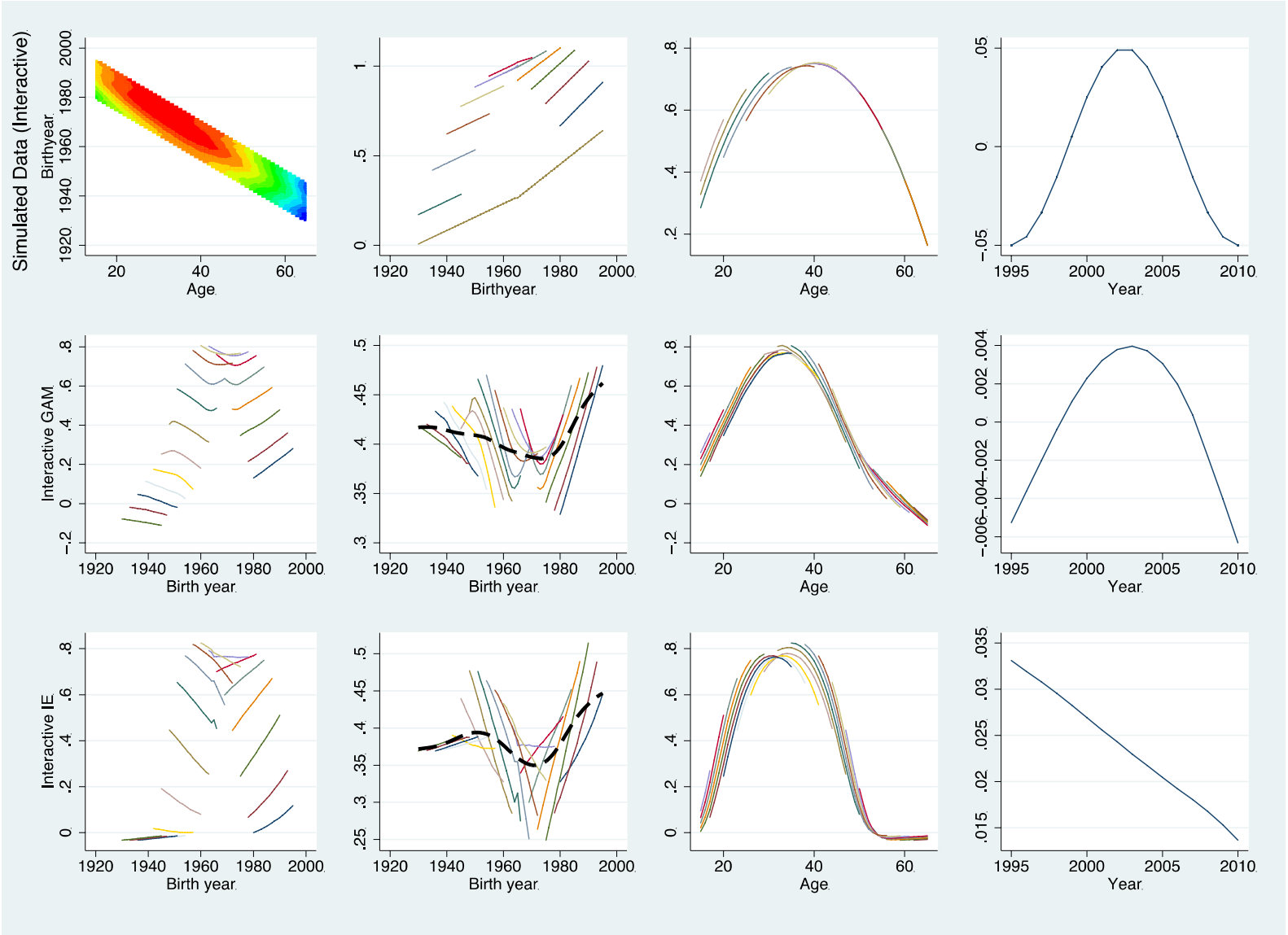

Figure 4.4 Mean APC profiles from interactive GAM and IE models using simulated data

4.4. Application to black male South African labour force participation

This section therefore compares the performance of additive and interactive models with real data, to assess to which extent it is necessary to impose the assumption of non-permanent cohort effects over the life cycle. In addition, the model selection procedures of McKenzie (2006) are illustrated to establish whether the known methods satisfy more general criteria. The benefit of applying the latter approach is that some economic knowledge can support the selection of statistical criteria in a very practical manner, a synergy which does not underlie any of the other methods discussed or applied in this study.

The data analyzes the evolution of black male labour force participation in South Africa. As noted above, the characteristic feature of realized labour force participation is greater levels of entry at young ages for more recent generations. The top left panel of Figure 4.7 shows this clearly. Between the ages of 18 and 35 (approximately), the raw data reveals that age profiles of recent labour market entrants are elevated. Figure 4.5 displays the same pattern in the form of a bivariate heat map, whereby it is clear that younger cohorts have higher participation rates early in the life cycle. However, this pattern disappears at later ages. Hence, the assumption of a permanent cohort effect is potentially unfounded, so that additive models can be misleading. The simulation evidence suggests that additive models may over-emphasize permanent cohort effects in such a situation.

Annual cross sectional data from the 1995 to 1999 October Household Survey and the September rounds of the 2000 to 2007 Labour Force Survey, which were collected by Statistics South Africa, are utilized to complete this analysis. In line with recommendations from earlier studies, the broad definition of labour force participation is applied, where discouraged workers are considered participants despite having ceased searching for a job (Kingdon & Knight, 2006b). The searching unemployed and the gainfully employed are also included in this definition. Only participants between the ages of 15 and 65 are investigated across the waves, in line with International Labour Organisation definitions. Cells for each of the *apc* observations are created in each of the waves, constructing a pseudo-panel dataset of appropriately weighted labour force participation rates for black males. This dataset satisfies the requirements of being treated as a true panel dataset, in that the group that is studied contains a sufficient number of underlying observations in each cell (Verbeek & Nijman, 1992; Burger & von Fintel, 2014). However, as Jiang & Carriere (2014) show, even settings with sparser data can be potentially well-estimated using GAMs. The simulation study presented above shows that even this condition does not ensure that estimates are reliable at endpoints in the data. Previous work has used additive APC models using only the Deaton-Paxon restriction to compare the evolution of labour force participation across demographic

groups (Burger & von Fintel, 2014); the current contribution focuses on one of these groups (black males) to understand whether the strong cohort effects previously found can be classified as permanent or temporary.

Apart from the models that were applied in the simulation study, McKenzie's (2006) approach for testable additive modelling is also implemented here. While it is possible to "automate" this procedure for simulation purposes, the discussion that follows will illustrate why this is not advisable: firstly, some breakpoints that are economically sensible are only marginally rejected statistically; secondly, some statistically significant changes cannot necessarily be matched directly with economic occurrences. Hence, the discretion of the researcher is required to choose optimal breaking points, combining statistical and economic knowledge.

First of all the additive intrinsic estimator is used as an initial estimate from which tests can be conducted. Given that all other estimators that were found to be reliable in the simulation study are a rotation of this estimator, and the curvature of profiles is identical for these approaches, the analysis is invariant to this choice (McKenzie, 2006). Figure 4.6 shows results from Wald tests at each point in the respective age, period and cohort profiles. Red lines indicate a p-value of 0.1, a potential threshold by which to discern a structural break in each of the components. Using this criterion, the cohort profile only changes for those born after 1948. It is not certain exactly why this is an influential breakpoint, as this group would have entered the labour market in the early 1960s, when a wide range of socio-political factors could have changed labour force participation patterns. Had the threshold for the p-value been set higher at 0.12, the 1983 birth cohort would also be classed at a structural break. In this case, the reasoning for the change (while only marginally statistically significant), can be motivated economically. This cohort was 15 years old at the implementation of school policies that restricted over-age learners from remaining in school, thereby changing participation behaviour (Burger et al., 2014). For this reason, it was decided to use a p-value threshold of 0.15 to discern structural breaks. By this criterion the birth cohort of 1942 also signifies a structural break.

At the same p-value, the age profile experiences breaks at 20. At the top end of the age distribution frequent changes occur at every age between 58 and 62, and again 64 to 65. Each of these ages should therefore be modelled using dummy variables. These observations indicate that the age profile is not perfectly quadratic over the entire life cycle: entry into the labour market and exit are both characterized by more sudden movements than over the rest of the life cycle. In terms of period effects, the tests identify multiple structural breaks around

the year 1999, when the business cycle turned into an upward phase. Hence, this test concurs with economic regularities.

Each of the identified segments is then tested in blocks, to establish whether linear or quadratic trends can be fitted in those regions of the data. Table 4.1 summarizes the fit that was found to be appropriate using Wald tests. In most cases both linearity and quadratic curves were rejected for all profiles. The most economically intuitive result emerging from this analysis is a linear trend after the 1999 turn in the business cycle, indicating an added worker effect. One could potentially proceed with only this restriction on the model, allowing *all* other coefficients to be estimated by dummy variables. Interestingly, the plausible break in the cohort profile at 1983 bears no consequence for estimation, as both segments around it are designated an exhaustive set of dummies. However, the analysis proceeds with each of the identified restrictions, despite only one necessary assumption to identify the model.

Table 4.1 Classification of APC profiles into linear, quadratic and flexible segments using Wald tests

Cohort			Age			Period		
Min	Max	Type	Min	Max	Type	Min	Max	Type
1930	1941	Dummies	15	19	Quadratic	1995	1996	Dummies
1942	1947	Linear	20	57	Dummies	1997	1999	Dummies
1948	1982	Dummies	58	64	Dummies	2000	2007	Linear
1983	1992	Dummies	65	65	Dummies			

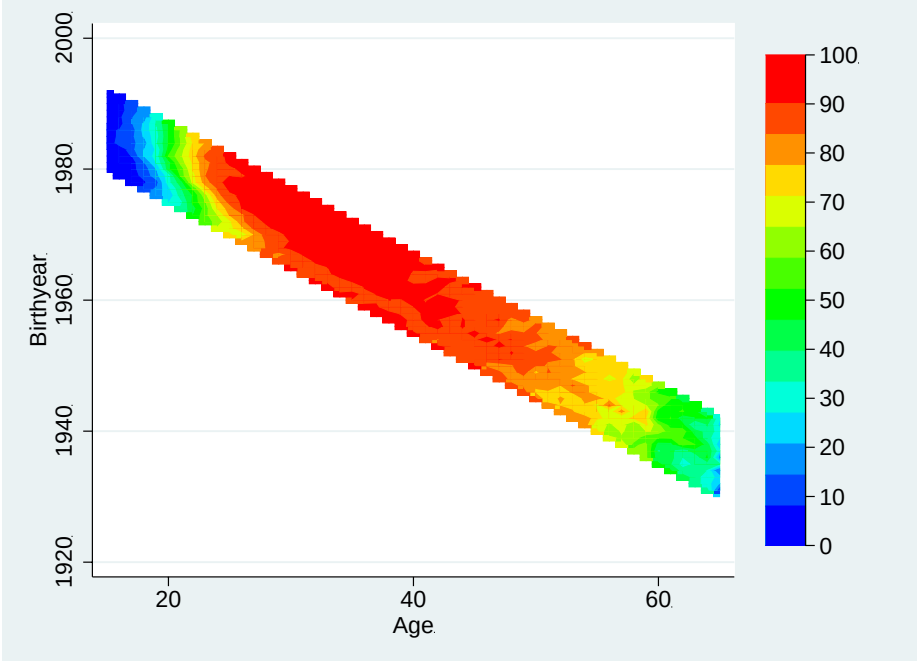

Figure 4.5 Labour force participation by age and birth year

NOTES: Own calculations from October Household Survey 1995-1999 and Labour Force Survey 2000-2007

Figure 4.7 summarizes the results from each of the additive models on the real data. As alluded to before, the top left panel confirms the pattern of earlier labour force entry for younger generations. As suggested by the simulation evidence, additive models could potentially model this as a "permanent" cohort feature, which will continue as a "formative experience" of this group, leaving them with a greater propensity to stay in the labour market later in life. Given that only 12 years of data is observed, it is not possible to conclude on the full life cycle behaviour for these generations. Hence, the permanence of cohort effects cannot be fully assessed. However, based on the fact that profiles do tend to assume a narrower band from roughly the age of 35, it is potentially true that these groups' high initial rates of labour market entry would slow down, so that by age 40 levels of participation could be the same for all generations. This would invalidate the assumption of a permanent cohort effect. Such a pattern is, however, consistent with a policy that affects a specific birth cohort at a specific point in their life cycle, such as the school policies analysed by Burger et al. (2014).

The age profiles in the top right panel of Figure 4.7 do not take this type of interaction into account. As confirmed by the simulation study, the Deaton-Paxon restriction and the maximum entropy approach appear to be rotations of the intrinsic estimator. The same is true for the McKenzie approach, where the first section of the age profile is fitted by a quadratic function. The two-phase semi-parametric approach also yields similarly-shaped estimates. The GAM age profile is more smoothly estimated, and places less emphasis on labour force participation early in the life cycle, while this is compensated for by high levels of participation in the cohort profile. Other models, except for the Deaton-Paxon and McKenzie restrictions, emphasize high participation rates early in a typical life cycle, affording less weight to the generational profiles. Hence, the cohort profiles from most models place less weight on younger *generations*, while using the GAM estimates the "average" life cycle to correspond to lower rates of participation *early in life*. The Deaton-Paxon and McKenzie restrictions emphasize both high early life cycle and young generation components, but have muted period effects. Many of these observations held true in the simulation study, where an interactive age profile was absorbed by cohort effects and where period effects were small.

Cohort effects in the bottom left panel assume that specific behaviour follow these groups over their lifetimes: the main feature identified by Burger & von Fintel (2014) is the distinct behaviour of generations born between 1975 and 1985 using the Deaton-Paxon restriction. While a gradual increase in cohort participation is consistent with greater inclusion of the black population into the labour market, the discontinuity in the data can be traced back to school policies that affected this group. The intrinsic estimator also matches this story to a large degree, though the increase in labour force participation is less stark. While the

maximum entropy profile is again a rotation of the intrinsic estimator, it estimates a downward cohort trajectory (nevertheless with a distinctive surge in participation for the 1975-1985 generation); it is apparent that it has placed more weight on the cohort effect for older generations, with a muted age profile at the end of the life cycle, so that these effects are confounded. These results are again similar to those found in the simulation study.

As with the simulation study, the semi-parametric estimator estimates a similar pattern to the intrinsic estimator; however, the magnitudes are again small relative to the other methods, so that even the distinct generational pattern that other methods discern seems to be negligible on the scale presented here. Turning to the GAM, the cohort profile consists of a long-run upward trend, reaching its peak just after 1975, after which it flattens off. One might consider this to be a problem of estimation in the tails, as younger cohorts are constructed from sparser data. However, as noted above, GAMs are supposedly good at dealing with this problem in the APC context (Jiang & Carriere, 2014). The simulation study contests this assertion, revealing that greater variability exists at profile endpoints (Figure E5). Rather, the GAM emphasizes a slower exit from the labour market at later *ages*, with low (permanent) participation rates for older *generations*. Though operating in opposite manners, both the GAM and the maximum entropy estimators highlight the dilemma inherent to the additive models: without interactive effects, the age and cohort effects are confounded.

The period effects in the bottom right panel of Figure 4.7 deliver two distinct patterns. The first is a roughly anti-cyclical labour force participation pattern discerned by the Deaton-Paxon restriction, the McKenzie approach and the GAM, with a turning point close to the year 2000, or two years after that. Given that the South African business cycle turned from a downward to an upward phase at the end of 1999, the estimates reveal economically expected behaviour, which could potentially provide justification for a zero restriction on period effects (Deaton, 1997), and which the McKenzie tests revealed explicitly. The second pattern (evident in the IE, maximum entropy and semi-parametric estimates) appears to estimate a long-run trend, whereby there has been an aggregate increase in labour force participation over time for black males – it assumes, however, that this was not driven by particular cohorts or age groups as the other estimates reveal. In the case of the semi-parametric estimator, this trend is not likely to be credibly estimated: because the first stage of this procedure only estimates a two-way model accounting for age and period effects, it does not condition on cohort variation (which is modelled separately in the second step). As a result, these estimates do not identify cohort variation from period estimates effectively, and would therefore deliver a long-run trend. In the cases of the IE and maximum entropy estimators, however, the partial effects of each period are estimated simultaneously. While the preceding discussion noted that the cohort and age profiles were traded off, it is also clear that the cohort and period

profiles are emphasized differently. In the case of the cohort profiles, it suggests a long-run decline in participation for newer entrants – the only reason they joined the labour market in the numbers that they did is because there has been an aggregate increase in participation over time for *all* cohorts, including them. However, as the simulation study revealed, the IE and maximum entropy estimators are particularly bad at identifying turning points in period effects, so that these results are unlikely to be true.

The similarity between the Deaton-Paxson and McKenzie estimators is noted here, confirming that previous work has obtained results that are believable based on testable assumptions (Burger & von Fintel, 2014; Burger et al., 2014). Hence, if it is appropriate to decompose labour force participation in South Africa additively, it is quite certain that the rise in participation is a cohort trend, and not an aggregate rise for all groups over time. The implication is that policies should accommodate this *generation* in the case of low absorption, rather than the population as a whole. However, these estimates cannot discern whether this generation is affected over its entire life cycle or only temporarily, and whether this group requires permanent intervention. Hence, this pattern requires further investigation in interactive models.

In summary, then, additive models are sensitive to their identifying assumptions. APC effects are likely to remain confounded, especially if cohort effects are not really permanent across the life cycle. Hence, it is necessary to pursue a more flexible approach by adding interactions to the model. It is, however, useful to note that under the assumption of additivity, restrictions are testable using the approach of McKenzie (2006). In this case it is possible to confirm that labour force participation does follow an anti-cyclical pattern, with a surge for the youngest cohorts. While this evidence is backed by multiple estimates, it is necessary to distinguish this effect as either part of the permanent long-run upward generational trend, or as a phenomenon that is temporary and can subside as these cohorts age. Each of the assumption-agnostic approaches does not identify these effects in the same manner. This illustration emphasizes the need for incorporating economically justifiable restrictions, but also those that satisfy statistical tests. The combination of these criteria can successfully uncover APC profiles.

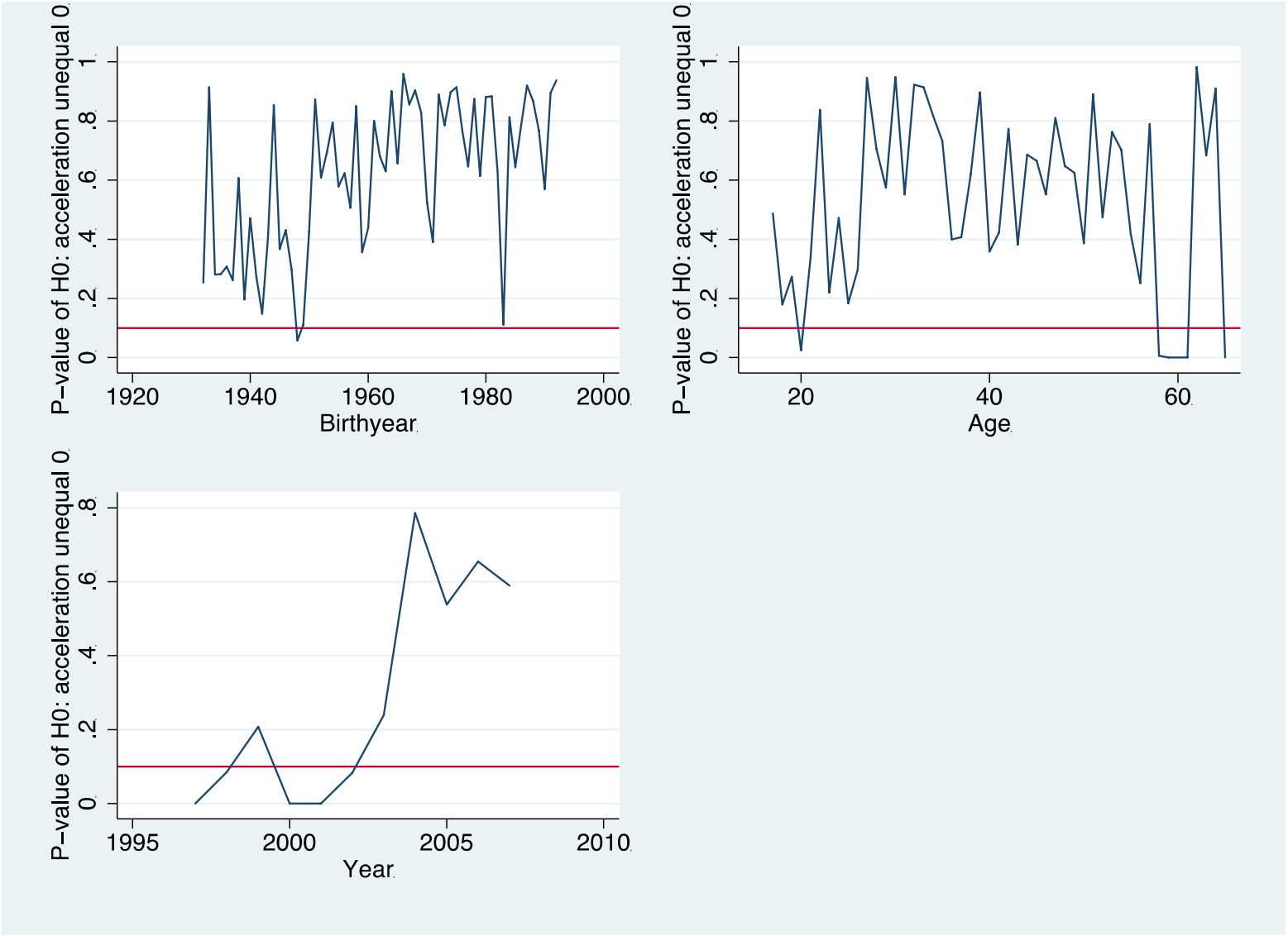

Figure 4.6 P-values of Wald tests to identify structural breaks using McKenzie (2006) approach

NOTES: Own calculations from October Household Survey 1995-1999 and Labour Force Survey 2000-2007

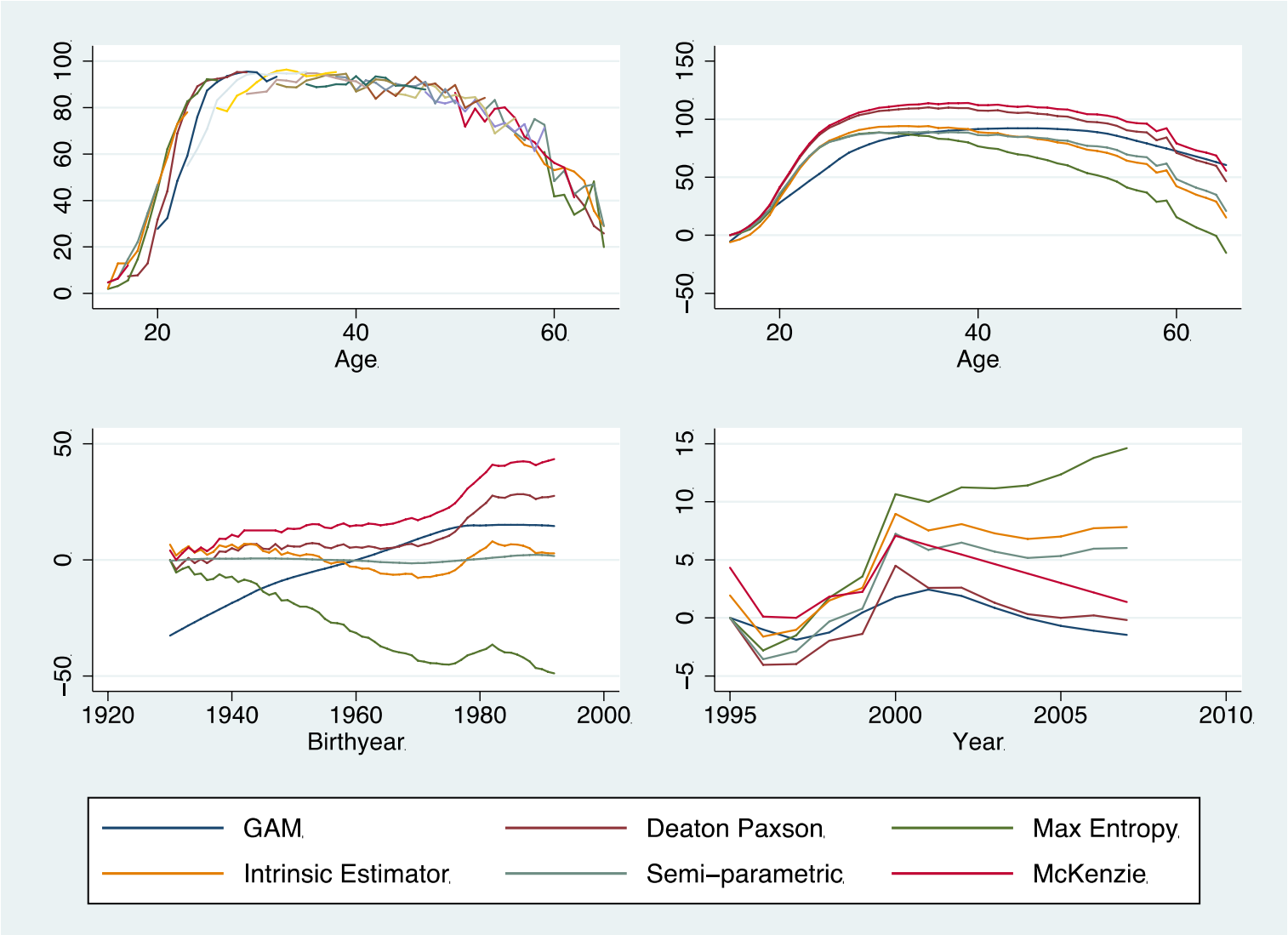

Figure 4.7 Additive APC models of South African labour force participation.

NOTES: Source: Own calculations from October Household Surveys (1995 to 1999) and September rounds of Labour Force Surveys (2000-2007). Top left panel represents raw data by birth cohort, other panels represent estimates from models.

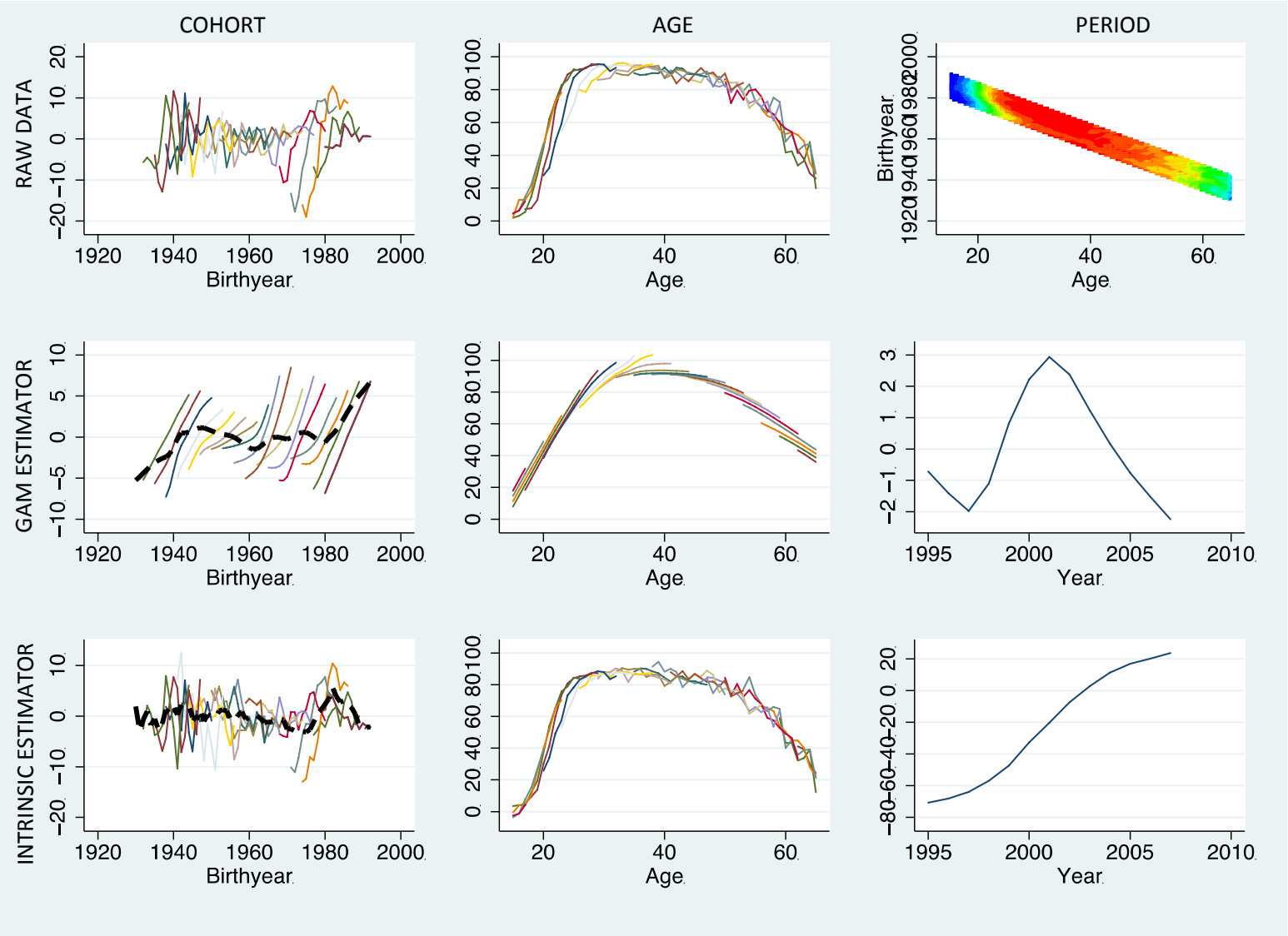

Figure 4.8 Interactive GAM and IE– APC profiles of South African labour force participation.

NOTES: Source: Own calculations from October Household Surveys (1995 to 1999) and September rounds of Labour Force Surveys (2000-2007)

Figure 4.8 extends the model to incorporate age and cohort interactions, with a separate period effect that does not vary by the other APC components. Both an interactive GAM and an interactive IE model are implemented. The first row depicts raw participation rates, while the second applies the interactive GAM and the third the interactive IE estimator. The first column represents cohort effects, presenting a trajectory for every third age group, and removing the average age effect for that group. Because the models are specified interactively, these cohort profiles are unique for each age group. However, to understand the overall effect, the dashed black line represents the average of each of these. The second column presents age effects differentiated by cohort, while the third considers period effects. In the latter case, the graph for the raw data provides a visual summary of participation, with the diagonal movements representing period effects, horizontal movements age effects, and vertical movements cohort effects.

The centre panel in the first row represents (as in Figure 4.7) the life cycle participation rate in the raw data of various birth cohorts, while the left panel in this row recasts this data by depicting cohort trajectories in the raw data, conditional on age and removing average age effects. The bottom panels repeat the same profiles, but this time trajectories are estimated from the interactive GAM and IE models. The most notable feature of the age profiles estimated by the GAM is that they are not as steep as the raw data would suggest at young ages. As in the simulated data, fanning occurs close to the turning point. While this is representative of greater entry into the labour market at younger ages for more recent generations, the occurrence appears at the "wrong age" using the interactive GAM, in much the same way as the simulation study suggested. Essentially, then, the GAM estimator introduces a time shift, whereby the earlier entry for younger generations occurs at slightly later ages. It is likely that the age effects are confounded with the other components, especially since the interaction between age and birth year is closely proxied by year effects, which are accurately determined by this method. The intrinsic estimator's age profile performs better on both counts, where the shape of the raw data is closely replicated by the predictions from the model: both the speed of the initial entry into the labour market, as well as the subsequent flattening off is well-captured by the estimates. The interactive IE performs better on this count in the real data.

While the GAM provides very smooth estimates for cohort trajectories, they tend to yield approximately parallel linear paths, falsely suggesting that an additive model would be sufficient to capture the necessary effects. The intrinsic estimator, on the other hand, emphasizes that some of the cohort effects flatten off, especially for the most recent generations. This is also visible in the raw data. As a result, it is in fact necessary to model this particular process interactively. In particular, the cohort effects for those born between

1975 and 1983 rise distinctly before flattening off, pointing to the fact that this is a temporary effect. The break identified by the Wald tests above reflects this pattern. The benefit of this estimator is that it has very clearly identified the very specific cohort effects, and their changes over the life cycle. Average profiles indicate that the interactive GAM simply captures an upward cohort effect, while the interactive IE discerns a particular surge for the youngest generations that resembles its additive counterpart. Again, the GAM tends to provide estimates that are different from other estimates at sample endpoints, indicating that corner observations are influential.

For the interactive GAM, period effects are very similar to those estimated in the additive GAM model: effects are small, but recover the anti-cyclical pattern effectively. As noted in the simulation study, the small magnitude of these estimates can potentially manifest in "time shifts" elsewhere in the model – in this case age profiles have been poorly estimated. Additionally the abrupt change of the period effect in 1996 concurs with the notion that GAM models do not perform well at endpoints, even when the data is not sparsely populated. Effects are much larger using the interactive IE estimator, but it is again apparent that this strategy is poor at uncovering cyclical variation; in this case, the model has only captured the upward phase, with no turning point. However, the model has estimated a much flatter cohort trajectory, so that the notion of a trade-off in interactive modelling is confirmed here.

The interactive approach followed in Figure 4.8 produces highly heterogeneous age and cohort effects, so that they are potentially difficult to generalize; in fact, extracting such patterns from the raw data (as in the top panels) would have produced similar conclusions, though without conditioning on period. Figure 4.9 provides one attempt at achieving more tractable results for interactive profiles. An interactive model is constructed based on the Wald tests shown in Figure 4.6 using the McKenzie (2006) approach. Importantly, the criteria are established assuming an additive model. However, cohort breakpoints can be exploited, so that a separate age profile can be fitted for each of the cohort categories. Furthermore, the linear and quadratic sections are adopted as outlined in Table 4.1. Given that the Deaton-Paxon results closely resemble this approach in the additive scenario (as in Figure 4.7.), the interactive distinction sheds light on whether earlier work captured a permanent or a temporary cohort effect (Burger & von Fintel, 2014).

Once the interactive age profiles are estimated by this method, the distinctive cohort surge for the youngest generations becomes somewhat muted. This is because the age profile of this group is slightly higher than that of the next group, though the difference is small. Apart from that, the age profiles are smooth and appear to "meet up", while the period effects are not different from an additive model. Comparing these estimates to the interactive IE, the specific

behaviour of the post-1975 generation has been de-emphasized in the cohort effects and shifted to the age profiles. The simulation evidence suggested that applying an additive model to a dataset that is really interactive over-emphasizes cohort effects for specific groups. Hence, the best evidence tends to be provided by profiles obtained from implementing testable assumptions. This is further corroborated by the ability of the latter to discern economic cycles, while the interactive IE is particularly weak at this.

The change in emphasis from permanent cohort effects to a temporarily different age profile clearly illustrates the concerns that additive APC models can mistakenly estimate permanent generational effects (Glenn, 1976). Rather, specific generations behave differently at only some points in their life cycle. The result is that we can expect the surge in participation (and the accompanying rise in unemployment) of younger generations to subside later in life, *ceteris paribus*. The results of Burger & von Fintel (2014), while correctly identifying a growing cohort effect, can be further broken down into a long-term linear upward trend and a short-run distinctive surge for the post-1975 generation that is not likely to continue into the future. However, it is necessary to observe these cohorts in later years to verify this implied path. The implication is that policies to help this generation out of unemployment may only need to be temporary in nature, and intervention later in their lives may not be necessary. The youth wage subsidy does exactly this, targeting individuals during a specific part of their life cycles (Levinsohn et al., 2014). However, if this cohort effect subsides, it is also possible that the problem of youth unemployment may be mitigated somewhat at a later time, as participation rates fall in line with the rest of the population.

The benefits evidenced by using testable assumptions in interactive decompositions indicate that it may be fruitful to investigate the possibility of multivariate Wald tests, in order to incorporate *directional* acceleration profiles, and finding two-way *surfaces*, which can be fitted with parametric hyperplanes. This approach is relegated to future work.

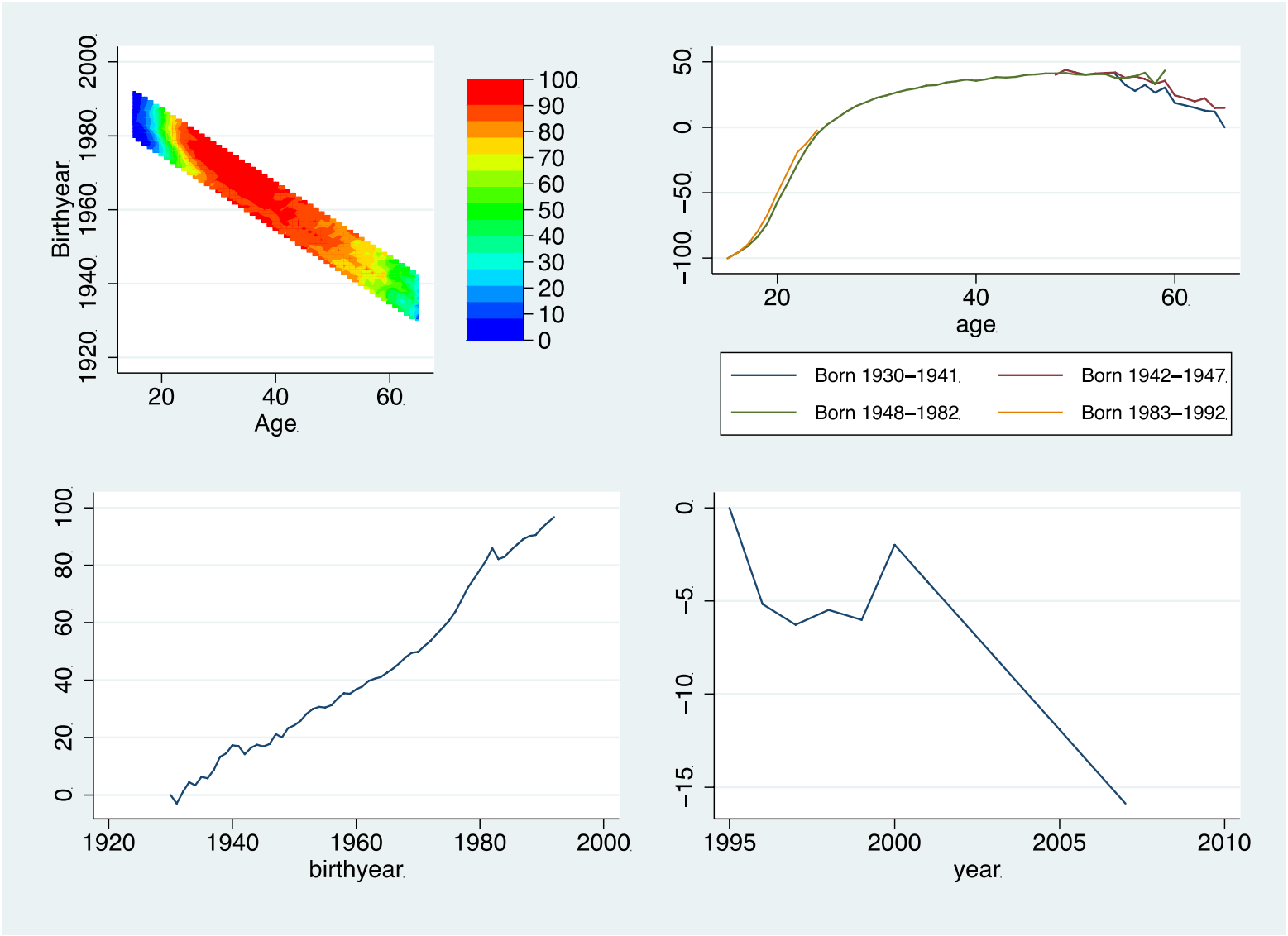

Figure 4.9 Interactive model of black male South African labour force participation using McKenzie breakpoints

NOTES: Source: Own calculations from October Household Surveys (1995 to 1999) and September rounds of Labour Force Surveys (2000-2007)

4.5. Conclusion

Rising labour force participation has lengthened the queue for jobs in South Africa, adding to a high unemployment rate. This long-run tendency is unlikely to reverse, with the emphasis shifting to devise policies that promote permanent absorption (Banerjee et al., 2008). How does evidence assist in targeting such interventions? Previous APC analysis has shown that a smooth upward generational trend in especially black labour force participation has emerged in South Africa, with distinctive behaviour for the post-1975 birth cohorts (Burger & von Fintel, 2014). Based on these results, high rates of participation among the current youth are interpreted as a generational effect that will continue across the life cycle of this generation. By this interpretation, high participation (and unemployment) rates are not specific to the "youth", but may only affect one generation. This *generation* should, supposedly, be targeted with policies that aid absorption into jobs, rather than youths at all points in time. However, the assumption that cohort effects are permanent may not be realistic, so that the distinctive surge in participation must be differentiated as either a pattern that has changed the long-run generational trajectory, or a temporary portion that may subside as these groups age. If the temporary effect dominates, it suggests that policy should rather target only a specific generation for a *part* of the life cycle. However, modifications to APC analysis are required to probe these possibilities.

While the APC decomposition has always had interpretative allure, its potential for mis-identification is large, especially if the data generating process is really interactive. A wide literature has documented the great sensitivity of estimates to (seemingly harmless) identification assumptions, even in the simpler additive setting. Recent contributions of other authors (Luo, 2013; Fienberg, 2013; Bell & Jones, 2014) re-iterate old concerns with the additive model (Glenn, 1976), based on new information and thorough evaluation. This study is the first (to the knowledge of the author) to relax the assumption that cohort profiles must be permanent.

Importantly, simulation evidence highlights that if one incorrectly assumes that the data generating is additive instead of interactive, cohort effects are amplified. This would lead to the false conclusion that temporary changes are permanent. IE and GAM models can be easily adapted to account for heterogeneous age profiles, and to separate cohort variation into permanent and temporary effects. IE models are popular in the current applied literature as an alternative when other commonly used behavioural assumptions cannot be justified. Nevertheless, this work emphasizes that the mechanics of this estimator do implicitly enforce behavioural assumptions on behalf of the researcher. Yet, it is not certain what these are. Furthermore, both Monte-Carlo simulations and applications to LFS data would suggest that the APC effects are confounded in both IE and GAM estimators. In particular, the

IE favours the estimation of long-run growth as an upward period trend rather than estimating it as a rising cohort effect, even when the true data generating process contains only cyclical time variation. Interactive GAM models are good at capturing cyclical shapes, but shift much of their magnitude to age effects. The latter are subsequently distorted. Furthermore, GAMs perform poorly at sample endpoints, questioning the results of Jiang & Carriere (2014).

However, most researchers have ignored the ability to statistically verify and implement behavioural assumptions in APC modelling (McKenzie, 2006). Following this approach confirms that previous estimates using the Deaton-Paxon (1994) restriction are largely valid, and identify economically consistent structural breaks, as well as profile slopes. It adequately identifies cyclical variation, and traces out the long-run generational increase in labour force participation. However, it is nevertheless possible that these additive models account for temporary cohort variations as permanent, as confirmed by the simulation study. Hence, the structural breaks are also implemented in an interactive setting. Cohort effects are dominated by a long-run *linear* cohort trend, with the distinctive surge amongst youngest generations captured by higher life cycle participation profiles early in these cohorts' careers. It furthermore appears that these distinct life cycle patterns tend to converge on those of older generations later in life, so that this higher participation amongst younger generations could potentially subside later in life. The McKenzie (2006) method is simple to implement, does not require researchers to guess the correct behavioural assumptions, nor does it impose unknown restrictions that other atheoretical approaches do. It furthermore provides results that are consistent with observations about the economy, but which are also statistically verified. Hence, it provides some answers to many outstanding questions in APC analysis, but more specifically provides a satisfactory description of labour market change in South Africa.

Effectively these estimates have revealed that a long-run increase in participation does affect younger generations (potentially permanently); however, the additional surge for the post-1975 birth cohort is likely to have only a temporary influence on the labour market. This behaviour is consistent with a temporary shock that was induced by policies that limited over-age learners from staying in the schooling system, and fuelling labour market entry amongst a young group of individuals (Burger et al., 2014). However, these cohorts would have entered the labour market in any event at later ages, so that the change in policy is unlikely to have a lasting effect. Using additional data to trace out the rest of these cohorts' life cycles will verify this assertion. These results suggest that this group requires additional policy assistance in finding jobs (but only early in life), over and above the concern of absorbing the surplus labour from younger and younger generations. Currently the youth wage subsidy that is being implemented focuses on young age groups. This appears to be correctly targeted as a measure that will not be available to young generations over their entire lives (if it is to account for the temporary surge amongst the youngest generations). However, it may not meet the demands of

rising *permanent* generational participation, unless its impact continues to a period beyond treatment. Future work should assess changing generational participation patterns in response to this policy.

Broadly this chapter has shown that it is possible to use APC analysis effectively using a combination of statistical and behavioural criteria, and that it is not impossible to relax the assumption of permanent cohort effects. It prompts other authors in all disciplines to continue with this type of analysis, despite earlier suggestions that this is a futile avenue of research (Fienberg, 2013; Glenn, 1976). More specifically, this chapter has also provided nuanced insight into generational change in South Africa's labour market, allowing for conjectures to be drawn about the permanence of current trends, and assisting policy discussion.

Chapter 5

Conclusions

South African labour market policy has benefited from more rigorous analysis by microeconometricians in the last two decades. This has been enabled by the availability of regularly enumerated large household surveys. The object of this dissertation has been to probe the reliability of a selection of these estimates, and to establish whether improved econometric solutions could provide a better understanding of labour market dynamics in South Africa. Whilst the labour market has evolved substantially even since before the fall of apartheid, many differences along spatial and racial lines persist. It is necessary to understand whether these rigidities are progressively being solved, or whether current policy continues to support them.

Three distinct econometric and policy scenarios have been investigated in the separate chapters. Chapter 2 highlights the importance of unobserved spatial heterogeneity in empirical modelling, especially where enforced spatial segregation from the past, and also institutional spatial rigidities (such as geographically defined bargaining agreements), directly influence labour market outcomes. Accounting for spatial heterogeneity dramatically changes estimates of labour market relationships. In particular, the approach separates a short-run from a long-run equilibrium in the context of wage curve estimation. The former is characterised by wage inflexibility, in contrast to past estimates that did not account for spatial heterogeneity (Kingdon & Knight, 2006a). Chapter 3 considered issues in the measurement of long-run mobility. In the infancy of cohort studies in South Africa, attempts at understanding the effects of childhood circumstances on adult labour market outcomes are limited. Alternative solutions in the form of recall data do exist, though they are compromised by anchored responses. Hence, data quality should necessarily improve to provide greater insight into long-run mobility within individuals' lifetimes, and whether childhood poverty transmits to labour market outcomes. Finally, Chapter 4 emphasized that in a changing political and social landscape, generational heterogeneity is influential in understanding labour market dynamics. The distinction between life cycle and generational tendencies has been highlighted, in order to show that specific events can affect some generations differently to others. However, based on assumptions of past analysis, results implied that behaviour would not change as birth cohorts aged; such conclusions would suggest that interventions target specific generations throughout their lives. In contrast, the current contribution suggests that generational heterogeneity may be non-permanent, and can dissipate over the life cycle. Various age-period-cohort estimators have been scrutinised in Chapter 4, concluding that a *non-permanent* surge in labour force participation occurred for recent cohorts, over and above a long-run permanent trend. While the separation of time variation into each of these components is complicated, this research has shown that it is wise to incorporate both economic and

statistical assumptions in modelling. It is not necessary to rely solely on statistical assumptions with their unknown economic implications (such as the use of the Intrinsic Estimator), nor is full knowledge of economic assumptions required. Under the assumption that profiles are truly additive, testable assumptions can be implemented (McKenzie, 2006). While this research has not presented new tests for the case that models are interactive, this type of heterogeneity can also be adequately captured by simple adaptations to these methods. As a result, many of the generational effects captured by previous work are found to be non-permanent.

The rest of this concluding chapter will elaborate on these findings, drawing together the main contributions of this dissertation and where relevant, highlight the potential for future research. Finally, the last section provides some closing remarks.

5.1. Spatial heterogeneity and local labour market definitions

Previous evidence suggests that South African wage setters are sensitive to slack local labour market conditions, moderating demands and offers in response to the long queue of job searchers within their regions of residence (Kingdon & Knight, 2006a). However, this behaviour is not mirrored in macroeconomic outcomes, with average real wages increasing over time despite economic downturns (Fedderke, 2012).

Re-investigation of these results in Chapter 2 confirms that wages are likely to be inflexible in response to local labour market conditions over the short-run. The negative relationship between wages and local unemployment that Kingdon & Knight (2006a) found is rather a reflection of a long-run spatial equilibrium. Without accounting for spatial heteregoneity, the long-run and short-run relationships are confounded. Data availability, however, limited previous studies from implementing appropriate estimation procedures. Time variation in micro data is required to include spatial fixed effects at the same geographic level at which unemployment is measured. These are influential in the estimation of South Africa's wage curve, in contrast to the findings of Nijkamp & Poot's(2005) meta-analysis of a sample of developed countries. The updated results in Chapter 2 agree more closely with those of countries where bargaining is highly centralized (Albaek et al., 2000; Daouli et al., 2013). Conclusions suggest that labour laws, in particular those relating to collective bargaining, are not conducive to the creation of employment, as other authors have also shown (Magruder, 2012; Rankin & Stijn, 2013). In addition, this study verifies that Blanchflower & Oswald (2008) are justified in their surprise that a country with an unemployment rate as high as South Africa's is wage flexible. Rather, their "empirical law of economics" that is embodied by the wage curve does not appear to be applicable in the short-run in an economy with high unemployment and high levels of centralized bargaining.

A second issue that is not explicitly explored by Nijkamp & Poot(2005) or any other authors in the wage curve literature, is the demarcation chosen to define a labour market. Many authors assume that the smaller the region, the closer it is to representing "local" conditions. However, in most cases researchers are constrained to choosing labour market definitions that are determined by surveys' sampling frames. This study is the first to highlight that wage curve estimates are biased if unemployment rates are calculated for regions that do not form truly functional labour market regions. Should demarcations be chosen that are either too small or too large, wage curve elasticities are biased towards zero. Using multiple demarcations in South Africa, Chapter 2 illustrates that district councils constitute more natural labour market boundaries than smaller magisterial districts and larger provinces, yielding the most negative point estimates. This concurs with the notion that collective bargaining agreements often cover multiple magisterial districts and galvanize large areas into functional labour market regions.

Additionally, Chapter 2 shows that the common use of time lagged unemployment rates as instruments in wage curve analysis is not effective in all instances. As South Africa's unemployment rate is structural, time variation is small. Once accounting for spatial heterogeneity, the permanent component of unemployment is removed in both first and second stages of instrumental variables estimates. The transitory component of unemployment is only weakly correlated over time. As an alternative, spatial lags of unemployment serve as instruments. In this case, instruments are strong. However, arguing for their exogeneity is more tenuous. Despite the problems with these instrumental variables, no estimate that controlled for spatial heterogeneity yielded a statistically significant wage curve effect. Hence, instrumentation does not increase wage flexibility as in other countries (Baltagi et al., 2012). Best estimates that account for spatial heterogeneity, measurement error and reverse causality indicate that a null effect prevails, rendering the labour market wage inflexible.

Future research can seek to better understand the direct influence of collective bargaining, by matching Labour Force Survey data with information on agreements that cover various industries and regions. Additionally, these estimates should ideally control for individual unobservables, for which pure panel data is required. It is also necessary to find an instrumental variable that is both strong and exogenous. These innovations will improve our understanding of how far reaching the impact of labour legislation is on the South African labour market.

5.2. Recalling childhood and long-run mobility

Chapter 3 has shown that while retrospective data can potentially add to our understanding of the effect of childhood background on adult outcomes, the recollection of childhood can be poor in developing countries. Using two waves of a nationally representative panel dataset allowed for the investigation of repeated self-evaluated rankings of individuals' childhood socioeconomic status and

the recall of parental attributes. Many adults seem willing to volunteer subjective rankings about their childhood socioeconomic status, though this kind of data is inconsistently reported by the same individuals over time. In particular, the analysis shows that the young are quick to adjust their rankings, while older individuals are more consistent in their answers. This suggests that length of recall since childhood is not influential in this type of retrospective reporting. Hence, other cognitive biases may be influencing individuals' recall.

To illustrate, comparisons are also made with recall of parental education. Individuals tend to have lower response rates on these indicators, despite the fact individuals are asked to recall the past on an objective scale. One reason is the large rate of parental absence in South Africa, owing to declining marriage rates and the migrant labour system (Posel & Rudwick, 2013; Posel, 2010). Individuals are therefore more likely to offer information about their own experiences, compared to information about their parents. However, those that do respond to these questions do so far more consistently across two waves of data. This suggests that objective questions are better recalled.

Why are subjective recall questions inconsistently answered? Two of the strongest correlates of adjusted childhood rankings are how individuals have changed their ratings of their position in the current village income distribution and their rating of subjective well-being across two waves. This is a clear sign of anchoring, whereby individuals first form a judgment about the present, which they "backcast" to update their view of the past. Importantly, changes in these perception measures are not correlated with changes in assessments of parental education. Hence, objective measures are less sensitive to anchoring effects, but unfortunately have low response rates.

The analysis of the influence of childhood on adult outcomes is compromised by anchoring. When a shock to the dependent variable (such as employment) also leads respondents to anchor subjective recall measures, reverse causality is introduced. Because repeated measures are available over time, a natural solution would be to instrument one assessment of childhood with another, as others have done in understanding the effects of health on employment (Crossley & Kennedy, 2002). However, if some errors in assessing the past (such as anchoring) are correlated over time, even this approach will lead to biased and inconsistent estimates of the effect of childhood on adult outcomes.

Because not one estimate can identify the true causal effect of childhood on adult (labour market) outcomes with the contaminated data, future work will concentrate on how a collection of inconsistent OLS and IV regressions can be used to construct bounds for this effect.

5.3. Heterogeneous age-period-cohort analyses

Separating life cycle from generational and time effects in any demographic variable has been a difficult task since social scientists started to pursue this line of empirical analysis many decades ago.

Whilst results are sensitive to behavioural identification assumptions (such as identical effects for adjacent cohorts, or a business cycle effect in time (Deaton, 1997)), other attempts have been devised to solve these types of problems atheoretically (Yang et al., 2004; Browning et al., 2012). However, even atheoretical methods implicitly make unknown behavioural assumptions, though it is not immediately clear what they are, as the statistical procedure imposes them automatically. Hence, there is no *a priori* choice of estimator, even among the class that bypasses the need to explicitly implement behavioural assumptions. McKenzie(2006), however, suggests a procedure by which assumptions are testable, allowing researchers to identify linear and quadratic sub-sections of any of the three profiles. Therefore, behavioural knowledge, together with statistical testing, can successfully trace out the age-period-cohort profiles, without having to guess about assumptions. Chapter 4 illustrated that this is a real possibility.

This contribution was the first to simulate a data generating process that had heterogeneous life cycle effects, specific to each cohort. This type of process enables cohort effects to be non-permanent, affecting these groups only during one part of their life cycle. Could any of the available additive estimators be extended to model heterogeneous age-period-cohort profiles successfully? Importantly, when the data generating process is truly interactive, all additive models generally over-emphasize permanent cohort effects, while in reality they could be temporary.

Interactive Generalized Additive Models (Hastie & Tibshirani, 1990) are a highly flexible way to do this. However, this potential solution, as in many other non-parametric settings, does not perform well at sample endpoints. Furthermore, this estimator is unique only when surfaces do not suffer from perfect concurvity, so that this is not a general solution to age-period-cohort problems. An adapted Intrinsic Estimator (Yang et al., 2004) that includes interactive profiles is presented for the first time. Simulation evidence suggests that this highly flexible estimation procedure is useful at recovering at least two of the profiles, even when cohort effects are temporary over the life cycle. Cyclical effects are the exception, with only one phase adequately reflected by the estimator. Application to LFS data reveals very similar results. Hence, these very flexible estimators still confound the separate effects of the three profiles.

However, it is necessary to find criteria to generalize results. The approach of McKenzie (2006) is useful in this regard. Using statistical criteria, breakpoints in each of the profiles were identified. Proceeding with estimates that modelled a specific life cycle effect for each of the distinct cohort groupings provides a more tractable solution to heterogeneous modelling. However, it may be necessary to also conduct the testing interactively, by instead finding linear or curved *surfaces* across two of the components. This approach has not yet been devised, and is relegated to future research.

How have heterogeneous solutions improved our knowledge of generational change in the South African labour market? All results are clear on the fact that recent generations entered the labour

market earlier than their predecessors. Is the gap in participation at young ages likely to persist over these cohorts' entire life cycles until retirement? The interactive solutions presented here indicate that this is not the case. Faster entry at the beginning of younger generations' working lives is offset by slower entry at later ages. As a result, by the middle of the life cycle, all generations seem to follow similar behaviour.

Given that these new generations also face high unemployment (because participation rates are not matched by the same strong levels of absorption), policies should focus on this particular group's earlier transition into jobs. It also suggests that such policies should focus their attention on individuals early in their life cycles, as these cohorts' behaviour starts to resemble that of older generations once they age. The youth wage subsidy represents one such policy that seems to target the correct cohort at an appropriate time in their lives (Levinsohn et al., 2014). Should this policy be applied permanently or should it be phased out after a certain period? While results presented here are not conclusive to answer this question, some preliminary thoughts are warranted. Previous work suggests that a large component of the surge in participation for this particular generation occurred because they were affected by a change in school policies that forced overage individuals into the labour market (Burger et al., 2014). Essentially this represents a policy failure, as the intended alternative for this group was to enter FET training colleges in order to become more employable. However, they did not take up this option. Initial indications suggest that the youth wage subsidy does have greater take-up. Later data will, however, be required to see whether the surge in labour force participation is permanent (or even increasing), and will give an indication of the required longevity of the youth wage subsidy.

5.4. Summary and final conclusions

This dissertation has tasked itself with finding improved microeconometric solutions to selected labour market questions in South Africa. Evidence for this country has, however, also contributed to methodological concerns in the broader literature.

It has shown that spatial heterogeneity in wage curve analysis is more important than previous authors have found to be the case. This is particularly true when highly centralized bargaining is prevalent in an economy. While instrumentation is vital in this estimation context, standard approaches fail in countries where unemployment is structural and persistent.

The dissertation has highlighted the need for better data on childhood circumstances in order to estimate life course mobility. While cohort data would be the first prize, the improvement of survey instruments that gather retrospective assessments would greatly aid the expansion of this literature in

developing countries. Even if anchoring does make this type of analysis difficult, proposals for future research suggest that it may be possible to use a collection of estimates to establish bounds for the extent of mobility over the lifetime.

While some authors have concluded that age-period-cohort analysis is a fruitless endeavour (Glenn, 1976; Fienberg, 2013), this dissertation has shown that this is perhaps an overly pessimistic view. Behavioural assumptions are in fact testable, and even provide guidance in estimating sensibly restricted interactive models. As a result, cohort effects need not be estimated to be permanent over the life cycle, and they provide insight into temporary deviations in behaviour for particular generations. This greatly assists in correctly targeting groups at the relevant time in their lives with policy interventions.

Each of these contributions has shown that existing approaches to microeconometric analysis in South Africa can contribute substantially to commentary about policy. However, in each case it is also clear that researchers have a role to play in refining their methods to provide more nuanced policy recommendations. Future work is likely to contribute further to the questions answered by this dissertation.

Reference List

Albaek, K., Asplund, R., Blomskog, S., Barth, E., Gumundsson, B.R., Karlsson, V. & Madsen, E.S., 2000. Dimensions of the wage-unemployment relationship in the nordic countries: wage flexibility without wage curve. In S.W. Polachek, ed. *Worker Well-being - Research in Labour Economics*. pp.345-81.

Almond, D. & Currie, J., 2011. Human Capital Development before Age Five. In D. Card & O. Ashenfelter, eds. *Handbook of Labor Economics*. Amsterdam: Elsevier. pp.1315-486.

Arbia, G., 2006. *Spatial Econometrics - Statistical Foundations and Applications to Regional Convergence*. Berlin: Springer.

Baltagi, B.H. & Blien, U., 1998. The German wage curve: evidence from the IAB employment sample.. *Economics Letters*, 61, pp.135-42.

Baltagi, B.H., Baskaya, Y.S. & Hulagu, T., 2012. The Turkish wage curve: Evidence from the Household Labor Force Survey. *Economics Letters*, 114(1), pp.128-31.

Baltagi, B.H., Blien, U. & Wolf, K., 2000. The East German wage curve 1993–1998. *Economics Letters*, 69, pp.25-31.

Banerjee, A., Galiani, S., Levinsohn, J., McLaren, Z. & Woolard, I., 2008. Why has unemployment risen in the New South Africa? *Economics of Transition*, 16(4), pp.715-40.

Batty, G.D., Lawlor, D.A., Macintyre, S., Clark, H. & Leon, D.A., 2005. Accuracy of adults' recall of childhood social class: findings from the Aberdeen children of the 1950s study. *Journal of Epidemiology and Community Health*, 59, pp.898-903.

Beckett, M., De Vanzo, J., Sastry, N., Panis, C. & Peterson, C., 2001. The Quality of Retrospective Data: An Examination of Long-Term Recall in a Developing Country. *The Journal of Human Resources*, 36(3), pp.593-625.

Bell, A. & Jones, K., 2014. Another 'futile quest'? A simulation study of Yang and Land's Hierarchical Age-Period-Cohort model. *Demographic Research*, 30(11), pp.333-60.

Berney, L. & Blane, D., 1997. Collecting retrospective data: accuracy of recall after 50 years judged against historical records. *Social Science Medicine*, 45, pp.1519-25.

Bhorat, H. & Hodge, J., 1999. Decomposing Shifts in Labour Demand in South Africa. *The South African Journal of Economics*, 67(3), pp.349-80.

Bhorat, H., Goga, S. & van der Westhuizen, C., 2012. Institutional wage effects: revisiting union and bargaining council wage premia in South Africa. *South African Journal of Economics*, 80(3), pp.400-14.

Blanchflower, D.G. & Oswald, A.J., 1990. The Wage Curve. *NBER Working Paper*, 3181. National Bureau of Economic Research.

Blanchflower, D.G. & Oswald, A.J., 1994. *The Wage Curve*. Cambridge: MIT Press.

Blanchflower, D.G. & Oswald, A.J., 2008. Wage Curve. In S.N. Durlauf & L.E. Blume, eds. *The New Palgrave Dictionary of Economics*. 2nd ed. Palgrave Macmillan.

Blien, U., Dauth, W., Schank, T. & Schnabel, C., 2013. The Institutional Context of an 'Empirical Law': TheWage Curve under Different Regimes of Collective Bargaining. *British Journal of Industrial Relations*, 51(1), pp.59-79.

Branson, N. & Wittenberg, M., 2007. The measurement of employment status in South Africa using cohort analysis, 1994-2004. *South African Journal of Economics*, 75(2), pp.313-26.

Branson, N., Ardington, C., Lam, D. & Leibbrandt, M., 2013. *Changes in education, employment and earnings in South Africa - a cohort analysis*. SALDRU Working Paper 105. Cape Town: Southern Africa Labour and Development Research Unit University of Cape Town.

Brown, M., 2012. *Assessing recall of early life circumstances: Evidence from the National Child Development Study*. CLS Working Paper 2012/3. Centre for Longitudinal Studies.

Brown, N., Rips, L.J. & Shevell, S.L., 1986. The subjective dates of natural events in very long-term memory. *Cognitive Psychology*, 17, pp.139-77.

Browning, M. & Bonke, J., 2009. *Allocation within the household: direct survey evidence*. Discussion Paper 429. Oxford: Oxford University.

Browning, M. & Lechene, V., 2003. Children and Demand: Direct and Non-Direct Effects. *Review of Economics of the Household*, 1, pp.9-31.

Browning, M., Crawford, I. & Knoef, M., 2012. *The age-period-cohort problem: set identification and point identification*. CEMMAP Working Paper CWP02/12. London: Institute for Fiscal Studies.

Buettner, T., 1999. The effect of unemployment, aggregate wages and spatial contiguity on local wages: an investigation with German district level data. *Papers in Regional Science*, 78, pp.47-67.

Bundy, C., 1979. *The rise and fall of the South African peasantry*. London: Heinemann.

Burger, R. & Yu, D., 2006. Wage trends in post-apartheid South Africa: Constructing an earnings series for South Africa from household survey data. *Labour Market Frontiers*, 8, pp.1-8.

Burger, R.P. & von Fintel, D.P., 2014. Rising unemployment in a growing economy: a business cycle, generational and life cycle persepective of post-transition South Africa's labour market. *Studies in Economics and Econometrics*, 38(1).

Burger, R.P., van der Berg, S. & von Fintel, D.P., 2014. The Unintended Consequences of Education Policies on South African Participation and Unemployment. *South African Journal of Economics*, forthcoming.

Card, D., 1995. The Wage Curve: A Review. *Journal of Economic Literature*, XXXIII(June), pp.785-99.

Casale, D. & Posel, D., 2002. The continued feminization of the labour force in South Africa: an analysis of recent data and trends. *The South African Journal of Economics*, 70(1), pp.156-84.

Casale, D., Desmond, C. & Richter, L., 2014. The association between stunting and psychosocial development among children of preschool age: a study using the South African Birth to 20 data. *Child: Care, Health and Development*, forthcoming.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pp.36-46.

Crossley, T.F. & Kennedy, S., 2002. The reliability of self-assessed health status. *Journal of Health Economics*, 21, pp.643-58.

Cunha, F., Heckman, J.J., Lochner, L. & Masterov, D.V., 2006. Interpreting the evidence on life cycle skill formation. In E.A. Hanushek & F. Welch, eds. *Handbook of the Economics of Education*. Amsterdam: Elsevier. pp.697-812.

Daouli, J., Demoussis, M., Giannakopoulos, N. & Laliotis, I., 2013. *The wage curve during the great depression in Greece*. Bonn, 22-23 August: 8th IZA/World Bank Conference on Employment and Development.

Deaton, A. & Paxon, C., 1994. Saving, growth and aging in Taiwan. In D.A. Wise, ed. *Studies in the Economics of Aging*. Chicago: University of Chicago Press. pp.331-62.

Deaton, A., 1997. *The analysis of household surveys: a microeconometric approach to development policy*. Baltimore: Johns Hopkins University Press.

Driffill, J., 2006. The Centralization of Wage Bargaining Revisited: What have we Learnt? *Journal of Common Market Studies*, 44(4), pp.731-56.

Duflo, E., 2003. Grandmothers and Granddaughters: Old-Age Pensions and Intrahousehold Allocation in South Africa. *The World Bank Economic Review*, 17(1), pp.1-25.

Easterlin, R.A., 2001. Income and happiness: towards a unified theory. *The Economic Journal*, 111, pp.465-84.

Fedderke, J., 2012. The Cost of Rigidity: the case of the South African labor market. *Comparative Economic Studies*, 54, pp.809-842.

Fenske, J. & Kala, N., 2012. *Climate, Ecosystem resilience and the Slave Trade*. Oxford: Centre for the Study of African Economies Working Paper 2012-23.

Fields, G., 2011. Labor market analysis for developing countries. *Labour Economics*, 18, pp.S16-22.

Fienberg, S.E., 2013. Cohort Analysis' Unholy Quest: A Discussion. *Demography*, 50, pp.1981-84.

Fourie, F.C.v.N., 2011. *The South African unemployment debate: thee worlds, three discourses*. SALDRU Working Paper 63. Cape Town: Southern African Labour and Development Research Unit.

Finn, A., Leibbrandt, M. & Levinsohn, J., 2014. Income mobility in a high-inequality society: Evidence from the first two waves of the National Income Dynamics Study. *Development Southern Africa*, 31(1), pp.16-30.

Friedman, M., 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press.

Glenn, N.D., 1976. Cohort Analysis' Futile Quest: Statistical Attempts to Separate Age, Period, and Cohort Effects. *American Sociological Review*, 41(5), pp.900-04.

Gregg, P. & Tominey, E., 2005. The wage scar from male youth unemployment. *Labour Economics*, 12, pp.487-509.

Grün, C., 2003. *Racial and Gender Wage Differentials in South Africa: What can Cohort Data tell?* Munich Discussion Paper No. 2003-21. Munich: Department of Economics, Munich University.

Haas, S.A., 2007. The long-term effects of poor childhood health: an assessment and application of retrospective reports. *Demography*, 44(1), pp.113-35.

Hannaford, M.J., Staub, M., Jones, J.M. & Bigg, G.R., 2014. Climate variability and societal dynamics in pre-colonial southern African history (AD 900-1840). *Environment and History*, forthcoming.

Harper, S., Lynch, J., Hsu, W., Everson, S., Hillemeier, M., Raghunathan, T.E., Salonen, J.T. & Kaplan, G.A., 2002. Life course socioeconomic conditions and adult psychosocial functioning. *International Journal of Epidemiology*, 31, pp.395-403.

Harris, J. & Todaro, M.P., 1970. Migration, Unemployment, and Development: A Two-Sector Analysis. *The American Economic Review*, 60(March), pp.126-42.

Hastie, T. & Tibshirani, R., 1990. *Generalized Additive Models*. London: Chapman and Hall.

Hill, C., Hosegood, V. & Newell, M-L., 2008. Children's care and living arrangements in a high HIV prevalence area in rural South Africa. *Vulnerable Children and Youth Studies*, 3(1), pp.65-77.

Hofmeyr, A., 2010. Social networks and ethnic niches: an econometric analysis of the manufacturing sector in South Africa. *South African Journal of Economics*, 78(1), pp.107-30.

Imbens, G. & Wooldridge, J., 2007. *Control Function and Related Methods*. National Bureau of Economic Research.

Iverson, T., 1998. Wage Bargaining, Central Bank Independence, and the Real Effects of Money. *International Organization*, 52(3), pp.469-504.

Janssens, S. & Konings, J., 1998. One more wage curve: the case of Belgium. *Economics Letters*, 60, pp.223-27.

Jiang, B. & Carriere, K.C., 2014. Age-period-cohort models using smoothing splines: a generalized additive model approach. *Statistics in Medicine*, 33, pp.595-606.

Johnson, R.A. & Wichern, D.W., 2002. *Applied Multivariate Statistical Analysis*. Fifth International Edition ed. Upper Saddle River, NJ: Pearson.

Kaus, W., 2013. Conspicuous consumption and "race": Evidence from South Africa. *Journal of Development Economics*, 100, pp.63-73.

Kennedy, S. & Borland, J., 2000. A wage curve for Australia? *Oxford Economic Papers*, 52, pp.774-803.

Keyes, K.M. & Li, G., 2010. A multiphase methods for estimating cohort effects in age-period contingency table data. *Annals of Epidemiology*, 20(10), pp.779-85.

Kingdon, G. & Knight, J., 2004. Unemployment in South Africa: The Nature of the Beast. *World Development*, 32(3), pp.391-408.

Kingdon, G. & Knight, J., 2006a. How Flexible Are Wages in Response to Local Unemployment in South Africa? *Industrial and Labour Relations Review*, 59(3), pp.471-95.

Kingdon, G. & Knight, J., 2006b. The measurement of unemployment when unemployment is high. *Labour Economics*, 13, pp.291-315.

Klasen, S. & Woolard, I., 2009. Surviving Unemployment Without State Support: Unemployment and Household Formation in South Africa. *Journal of African Economies*, 18(1), pp.1-51.

Klein, N., 2012. *Real Wage, Labor Productivity, and Employment Trends in South Africa: A Closer Look*. IMF Working Paper 12/92. Washington, DC: International Monetary Fund.

Kriegler, N., Okamoto, A. & Selby,J.V., 1998. Adult female twins' recall of childhood social class and father's education: a validation study for public health research. *American Journal of Epidemiology*, 147(7), pp.704-08.

Lechtenfeld, T. & Zoch, A., 2014. *Income Convergence in South Africa: Fact or Measurement Error?* Stellenbosch Economic Working Paper 10/14. Department of Economics, Stellenbosch University.

Leibbrandt, M., Woolard, C. & Woolard, I., 2009. Poverty and Inequality Dynamics in South Africa: Post-Apartheid Developments in the Light of the Long-Run Legacy. In J. Aron, B. Kahn & G. Kingdon, eds. *South African Economic Policy under Democracy*. Oxford. Oxford University Press.

Levinsohn, J., Rankin, N., Roberts, G. & Schoer, V., 2014. *Wage subsidies and youth employment in South Africa: Evidence from a randomised control trial*. Stellenbosch Economic Working Paper 02/14. University of Stellenbosch.

Lewis, J.D., 2001. World Bank Informal Discussion Paper on Aspects of the Economy of South Africa 16.

Luo, L., 2013. Assessing validity and application scope of the intrinsic estimator approach to the age-period-cohort problem. *Demography*, 50, pp.1945-67.

Madhavan, S., Richter, L., Norris, S. & Hosegood, V., 2014. Fathers' Financial Support of Children in a Low Income Community in South Africa. *Journal of Family Economics Issues*, forthcoming.

Magruder, J., 2012. High Unemployment Yet Few Small Firms: The Role of Centralized Bargaining in South Africa. *American Economic Journal: Applied Economics*, 4(3), pp.138-66.

Mariotti, M., 2012. Estimating the substitutability of African and White workers in South African manufacturing, 1950–1985. *Economic History of Developing Regions*, 27(2), pp.47-60.

Mason, K.O., Mason, W.M., Winsborough, H.H. & Poole, W.K., 1973. Some Methodological Issues in Cohort Analysis of Archival Data. *American Sociological Review*, 38(2), pp.242-58.

Matsuura, K. & Willmot, C.J., 2012. *Terrestrial Precipitation: 1900-2010 Gridded Monthly Time Series*. [Online] Available at: http://climate.geog.udel.edu/~climate/html_pages/Global2011/Precip_revised_3.02/README.GlobalTsP2011.html [Accessed 2014].

McKenzie, D., 2006. Disentangling age, cohort and time effects in the additive model. *Oxford Bulletin of Economics and Statistics*, 68(4), pp.473-95.

McKenzie, S.K. & Carter, K.N., 2009. Are retrospective measures of childhood socioeconomic position in prospective adult health surveys useful? *Australasian Epidemiologist*, 16(3), pp.22-24.

Mincer, J., 1962. On-the-job training: costs, returns and some implications. *Journal of Political Economy*, 70(5), pp.S50-79.

Mittelhammer, R.C., Judge, G.G. & Miller, D.J., 2000. *Econometric Foundations*. Cambridge: Cambridge University Press.

Moretti, E., 2011. Local Labor Markets. In D. Card & O. Ashenfelter, eds. *Handbook of Labor Economics*. Amsterdam: Elsevier. pp.1237-313.

Moulton, B.R., 1990. An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units. *The Review of Economics and Statistics*, 72(2), pp.334-38.

Nel, J.H., Krygsman, S.C. & de Jong, T., 2008. The identification of possible future provincial boundaries for South Africa based on an intramax analysis of journey-to-work data. *ORiON*, 24(2), pp.131-56.

Nijkamp, P. & Poot, J., 2005. The Last Word on the Wage Curve? *Journal of Economic Surveys*, 19(3), pp.421-50.

Nunn, N. & Pugo, D., 2012. Ruggedness: The blessing of bad geography in Africa. *Review of Economics and Statistics*, 91(1), pp.20-36.

Pannenberg, M. & Schwarze, J., 1998. Labor market slack and the wage curve. *Economics Letters*, 58, pp.351-54.

Papps, K.L., 2001. Investigating a Wage Curve for New Zealand. *New Zealand Economic Papers*, 35(2), pp.218-39.

Pienaar, L. & von Fintel, D.P., 2014. Hunger in the former apartheid homelands: determinants of convergence one century after the 1913 Land Act. *Agrekon*, Forthcoming.

Piraino, P., 2014. *Intergenerational earnings mobility and equality of opportunity in South Africa*. SALDRU Working Paper 131. Cape Town: Southern African Labour and Development Research Unit.

Posel, D. & Rudwick, S., 2013. Changing patterns of marriage and cohabitation in South Africa. *Acta Juridica*, 13(1), pp.169-80.

Posel, D., 2010. Households and Labour Migration in Post-Apartheid South Africa. *Studies in Economics and Econometrics*, 34(3), pp.129-41.

Rankin, N. & Stijn, J.-P., 2013. *Productivity, wages and technology choice in South African manufacturing firms*. Mimeograph. Stellenbosch: Stellenbosch University.

Republic of South Africa, 2012. *National Development Plan 2030: Our Future - make it work*. Pretoria: National Planning Commission.

Roback, J., 1982. Wages, rents and the quality of life. *Journal of Political Economy*, 90, pp.1257-78.

Rosen, S., 1979. Wagebased indexes of urban quality of life. In P.N. Miezkowski & M.R. Straszheim, eds. *Current Issues in Urban Economics*. Baltimore, MD: Johns Hopkins University Press. pp.74-104.

Rosenzweig, M.R. & Udry, C., 2014. *Rainfall forecasts, weather and wages over the agricultural production cycle*. Cambridge, MA: National Bureau for Economic Research Working Paper 19808.

Sanz-de-Galdeano, A. & Turunen, J., 2006. The euro area wage curve. *Economics Letters*, 92, pp.93-98.

Schoer, V., Rankin, N. & Roberts, G., 2014. Accessing the first job in a slack labour market: job matching in South Africa. *Journal of International Development*, 26, pp.1-22.

Shannon, C.E., 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, pp.379-423.

Simini, F., Gonzalez, M.C., Maritan, A. & Barabasi, A.-L., 2012. A universal model for mobility and migration patterns. *Nature*, 484, pp.96-100.

Straughen, J.K., Caldwell, C.H., Osypuk, T.L., Helmkamp, L. & Misra, D.P., 2013. Direct and proxy recall of childhood socio-economic position and health. *Paediatric and Perinatal Epidemiology*, 27, pp.294-302.

Tukey, J.W., 1977. *Exploratory Data Analysis*. Reading, MS: Addison-Wesley Publishing Company.

Tversky, A. & Kahneman, D., 1974. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), pp.1124-31.

Union of South Africa, 1912. *Census of the Union of South Africa 1911*. Pretoria: Government Printer.

Verbeek, M. & Nijman, T., 1992. Can cohort data be treated as genuine panel data? *Empirical Economics*, 17, pp.9-23.

Wagner, J., 1994. German Wage Curves, 1979-1990. *Economics Letters*, 44, pp.307-11.

Ward, M., 2011. Concordance of siblings' recall of measures of childhood socioeconomic position. *BMC Medical Research Methodology*, 11, p.147.

Winter-Ebner, R., 1996. Wage curve, unemployment duration and compensating differentials. *Labour Economics*, 3, pp.425-34.

Wooldridge, J.M., 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

World Economic Forum, 2014. *The Global Competitiveness Report 2013-2014*. World Economic Forum.

Yang, Y. & Land, K.C., 2006. A mixed models approach to the age-period-cohort analysis of repeated cross section surveys, with an application to data on trends in verbal test scores. *Sociological Methodology*, 36(1), pp.75-97.

Yang, Y. & Land, K.C., 2013. Misunderstandings, Mischaracterizations, and the Problematic Choice of a Specific Instance in Which the IE Should Never Be Applied. *Demography*, 50, pp.1969-71.

Yang, Y., Fu, W.J. & Land, K.C., 2004. A methodological comparison of age-period-cohort models: the intrinsic estimator and conventional generalized linear models. *Sociological Methodology*, 34(1), pp.75-110.

Yang, Y., Schulhofer-Wohl, S., Fu, W.J. & Land, K.C., 2008. The intrinsic estimator for age-period-cohort analysis: What it is and how to use it. *American Journal of Sociology*, 113(6), pp.1697-936.

Appendix A
Short-run inflexibility by industrial structure

Given the evidence of Magruder (2012), the concurrence of missing small formal sector firms, high wages and lower employment can be explained by collective bargaining, as also postulated by Rankin & Stijn (2013). This estimation context further investigates the wage inflexibility that was found in Chapter 2, by splitting the sample to understand whether rigidities can be traced to a bargaining channel.

Kingdon & Knight (2006a), for instance, find that the group of union members in South Africa does not exhibit wage curve behaviour, while non-union members do. Blien et al. (2011) show that workers in Western Germany who bargain at a more centralized level, beyond their firm, are less flexible in their wage setting than those that bargain only at the firm level. These arguments are consistent with findings that countries with more centralized bargaining do not have wage curves (Albaek et al., 2000; Daouli et al., 2013).

Table A1 estimates separate wage curve elasticities by workers' unionisation status and by their firm size. Fixed effects estimates, using district council unemployment rates[32], are shown with 95% confidence intervals.

Considering the elasticities for union members and non-unionized workers (without conditioning on firm size) the results of previous studies are confirmed (Kingdon & Knight, 2006a; Blien et al., 2011). Union members are insensitive to local unemployment rates, with positive or statistically insignificant elasticities. These workers tend to raise wage demands despite high local unemployment. The dividend of economic growth over the period of analysis accrued to higher wages of the employed rather than higher levels of employment creation. Results for this group most closely resemble those of the population as a whole, as discussed in Chapter 2, so that wage inflexibility in the South African labour market is associated with union members. Point estimates for non-unionized workers are negative, except for Indian males and white females. While some elasticities are insignificant, their 95% confidence intervals include values well below the international norm of -0.1. Institutional bargaining therefore tends to reduce workers' sensitivity to local labour market conditions.

Non-unionized workers may still benefit from bargaining through the extension of wages negotiated by industry bargaining councils and large unions. Because it is not possible to identify whether a

[32] This demarcation is chosen, because Chapter 2 shows it to be the most optimal. Instrumentation is avoided here, as instruments are weak at the district council level.

worker is subject to a collective bargaining agreement in the data, the effects of this institution are discussed indirectly. Since large unions and large companies usually bargain collectively (Magruder, 2012), the combination of union membership and working in a large firm proxies for this classification. Regardless of union membership, individuals that work in small firms of 1 to 4 employees are sensitive to high local unemployment rates, exhibiting wage curve behaviour (except for unionized Indian males and non-unionized white females). For females, and white and Indian males the point estimates are positive, though their confidence intervals mostly include large negative values. As individuals tend to work in slightly larger firms (5-49 employees), workers' responsiveness to unemployment generally becomes smaller (though some exceptions exist). This is indicative of the prevalence of collective bargaining in these sections of industry. Even the non-unionized display lower sensitivity to regional unemployment rates as they work in larger firms, despite elasticities being muted in all firm size segments. Hence, wage determination is not identical across small firm and large firm sectors, with collective bargaining in the large firm sector diminishing wage flexibility. Furthermore, the influence of collective bargaining in the large firm sector also spills over to members who are not a part of a union, as these agreements are extended to entire industries. These results show that non-union members benefit from collective agreements, while unionized workers additionally gain from firm level negotiations, as illustrated by Bhorat et al. (2012). However, workers in the very largest firms do not display these same patterns, so that the result is not robust.

The sensitivity of small firm workers to local labour market conditions suggests that they are less likely to be covered by collective bargaining agreements, or that compliance in paying collectively negotiated wages is low. The inclusion of the informal sector in these results may explain why bargaining effects do not extend wage inflexibility to small firms, as it does in slightly larger firms. Informal firms are less likely to comply with collective bargaining agreements.

High downward wage flexibility in the small firm sector would also provide a mechanism to at least partially clear the labour market, so that this sector would be a net job creator. However, this contrasts with the conclusions of Magruder (2012), who suggests that small firms' job creation capacity is dimished by the extension of collective bargaining agreements to entire industries. His analysis considers industries that are not typically active in the informal sector. Consequently, the results can be reconciled by the potential that wage curve behaviour amongst small firms is likely to be representative of informal sector workers who do not benefit from bargaining agreements. This does not, however, provide insight into why the informal sector does not create more jobs in South Africa, despite flexible wages when unemployment is high.

Table A.1 Wage curve elasticities by union and firm size status

Union membership	Union members				Non-unionized workers			
Firm size	1-4	5-49	50 plus	All	1-4	5-49	50 plus	All
Black male	-0.062	0.000	-0.020	0.185	-0.135	0.020	-0.047	-0.098
	(-0.285 ; 0.160)	(-0.132 ; 0.131)	(-0.149 ; 0.110)	(0.026 ; 0.344)	(-0.275 ; 0.004)	(-0.110 ; 0.149)	(-0.226 ; 0.131)	(-0.211 ; 0.014)
Black female	0.166	0.101	-0.257	-0.033	-0.171	-0.141	-0.365	-0.167
	(-0.136 ; 0.467)	(-0.055 ; 0.256)	(-0.432 ; -0.083)	(-0.204 ; 0.139)	(-0.295 ; -0.047)	(-0.335 ; 0.053)	(-0.547 ; -0.183)	(-0.297 ; -0.038)
Coloured male	-0.146	-0.063	0.029	0.311	-0.158	0.109	0.252	-0.023
	(-0.462 ; 0.170)	(-0.277 ; 0.151)	(-0.100 ; 0.158)	(0.162 ; 0.460)	(-0.343 ; 0.027)	(-0.031 ; 0.250)	(0.115 ; 0.389)	(-0.133 ; 0.088)
Coloured female	0.071	0.129	0.048	0.279	-0.215	0.188	0.163	-0.033
	(-0.271 ; 0.413)	(-0.025 ; 0.282)	(-0.142 ; 0.239)	(0.124 ; 0.434)	(-0.416 ; -0.014)	(-0.001 ; 0.376)	(0.023 ; 0.302)	(-0.156 ; 0.090)
Indian male	0.992	0.200	0.179	0.668	0.308	0.110	-0.048	0.029
	(0.510 ; 1.474)	(-0.285 ; 0.686)	(-0.258 ; 0.616)	(0.365 ; 0.972)	(-0.258 ; 0.875)	(-0.393 ; 0.612)	(-0.604 ; 0.508)	(-0.286 ; 0.343)
Indian female	0.728	0.537	-0.146	0.555	0.345	0.090	-0.476	-0.049
	(-2.116 ; 3.572)	(0.158 ; 0.915)	(-0.726 ; 0.433)	(0.227 ; 0.883)	(-0.337 ; 1.028)	(-0.365 ; 0.546)	(-1.027 ; 0.075)	(-0.437 ; 0.338)
White male	0.244	0.093	-0.102	0.641	0.178	0.157	0.068	-0.011
	(-0.084 ; 0.573)	(-0.086 ; 0.272)	(-0.287 ; 0.082)	(0.411 ; 0.870)	(-0.110 ; 0.465)	(-0.035 ; 0.349)	(-0.157 ; 0.293)	(-0.150 ; 0.127)
White female	0.579	-0.011	-0.177	0.667	0.226	0.377	0.001	0.074
	(-0.090 ; 1.248)	(-0.402 ; 0.379)	(-0.384 ; 0.030)	(0.464 ; 0.870)	(0.063 ; 0.389)	(0.114 ; 0.640)	(-0.210 ; 0.212)	(-0.090 ; 0.238)

NOTES: 95% Confidence intervals in parentheses. Own calculations from LFS Sept 2000 to LFS March 2004. Drawn from a similar specification to Table 2.2, specification 4, but with race-firm size-union-log(unemployment) interactions added, on which these results are based. Monthly earnings measured at the individual level and deflated by a national CPI. Unemployment measured at the magisterial district level (354 in the data) according to the broad definition. Additional controls for *education and its square, age and its square, magisterial district education composition, magisterial district occupation composition, magisterial district sector composition, magisterial district union density, magisterial district firm size concentration*. Standard errors are clustered at the geographic level at which unemployment is measured.

Appendix B
Geographic and ethnic reach of labour markets

The results in Chapter 2 suggest that various groups respond in unique ways to overall labour market conditions, with the black population group exhibiting behaviour most closely related to the empirical norm in the literature. It is possible that groups other than the black population form distinct sub-labour markets, discarding information about labour market conditions relevant to the broader demographic. Furthermore, various geographic regions also played a role, with provinces being too large to be considered local. Hence, this appendix more clearly delineates the network and spatial effects that define labour markets. To do so, wage curves are estimated as before, but now separate unemployment rates by race are introduced. Do individuals regard labour market experiences of other groups *within* their region in forming wage decisions?

Table B.1 explores this by considering similar estimates to before, but estimating wage curves using broad magisterial district unemployment rates specific to various racial groups. The simplest specification (B1), without any fixed effects, reveals that black and white individuals are both most responsive to their own group's unemployment rates, where tests show that no own-group elasticity differs from -0.1. However, females from both races respond to the unemployment rate of the other group. The progressive introduction of various fixed effects in specifications B2 to B4 changes these results somewhat: a wage curve only exists for the black population, and only in relation to their own unemployment rates once magisterial district fixed effects are included. White groups do not respond to any unemployment rates, and no group responds to white unemployment rates in the normal fashion.

Hence, white individuals are not as sensitive to local labour market conditions (however defined) compared to the black group. The wage setting process is therefore distinct for this population, and is not dependent on whether unemployment is high, even within their own ethnic group. However, because unemployment rates are so distinctly low for whites, it is understandably difficult to discern a wage curve for individuals in small localities. Additionally, calculating unemployment rates for white individuals is more problematic, as fewer individuals are sampled within each magisterial district.

Table B.1 Wage curve elasticities by worker demographics (using race-specific magisterial district unemployment rates)

	B1	B2	B3	B4
log(black UR) x black male	-0.132*	-0.203***	-0.151***	-0.052
log(black UR) x black female	-0.159	-0.209**	-0.156**	-0.096
log(black UR) x white male	-0.005	-0.062	-0.006	0.055
log(black UR) x white female	-0.014	-0.061	-0.007	0.053
log(white UR) x black male	-0.029*	-0.014	0.004	0.018
log(white UR) x black female	-0.083**	-0.065**	-0.038**	-0.011
log(white UR) x white male	-0.066***	-0.046**	-0.025	-0.011
log(white UR) x white female	-0.081**	-0.065**	-0.045**	-0.029
Age	0.111***	0.111***	0.110***	0.107***
Age^2	-0.001***	-0.001***	-0.001***	-0.001***
Education	-0.018***	-0.021***	-0.023***	-0.026***
$Education^2$	0.010***	0.010***	0.009***	0.010***
Period FE	Y	Y	Y	Y
Race x Gender FE	Y	Y	Y	Y
MD Controls	N	N	N	N
Province FE	N	Y	N	N
DC FE	N	N	Y	N
MD FE	N	N	N	Y
Constant	3.952***	3.910***	4.293***	4.442***
R-squared	0.544	0.554	0.561	0.567
N	76724	76724	76724	76724
p-value for H_0: $\beta_{BM\ x\ black\ unempl} = -0.1$	0.647	0.054*	0.327	0.281
p-value for H_0: $\beta_{BF\ x\ black\ unempl} = -0.1$	0.617	0.209	0.365	0.949
p-value for H_0: $\beta_{WM\ x\ black\ unempl} = -0.1$	0.053*	0.490	0.080*	0.002***
p-value for H_0: $\beta_{WF\ x\ black\ unempl} = -0.1$	0.102	0.421	0.101	0.062*
p-value for H_0: $\beta_{BM\ x\ white\ unempl} = -0.1$	0.000***	0.000***	0.000***	0.000***
p-value for H_0: $\beta_{BF\ x\ white\ unempl} = -0.1$	0.630	0.194	0.000***	0.000***
p-value for H_0: $\beta_{WM\ x\ white\ unempl} = -0.1$	0.143	0.006***	0.000***	0.000***
p-value for H_0: $\beta_{WF\ x\ white\ unempl} = -0.1$	0.634	0.221	0.006***	0.000***

NOTES: * p<0.1, ** p<0.05, *** p<0.01. Own calculations from LFS 2000b to 2004a. Monthly earnings measured at the individual level and deflated by a national CPI. Race-specific unemployment measured at the magisterial district level (354 in the data) according to the broad definition. Additional controls for *magisterial district education composition, magisterial district occupation composition, magisterial district sector composition, magisterial district union density, magisterial district firm size concentration*. Standard errors are clustered at the geographic level at which unemployment is measured.

Other specifications explore the implementation of spatially lagged magisterial district unemployment rates. These composite indicators summarize information within specified radii from a specific point, weighted by the inverse of distances. However, they exclude information from the district of

residence, and therefore only measure information spillover from *other* regions, indicating how far geographically dispersed demographic networks stretch in the wage setting mechanism. Should the effect differ by the size of the radius (which determines the inclusion of regions in the calculation of spatially lagged unemployment), it provides a sense of how large labour markets are for various demographic groups. It is also possible to understand whether workers from various groups are substitutes or complements, in the sense that unemployment of other groups influences their wage setting.

Figure B.1 summarizes the results of wage curve elasticities across space and race. In the first panel, only controls for education and age are introduced. Black male and female wages are sensitive to high unemployment of their own group within 100km of the centroid of their own district (with elasticities close to -0.1). This relationship strengthens as information further afield is accounted for, up to a radius of about 1500km. This suggests that black individuals are incorporated into labour markets (within their own demographic network) that stretch far beyond their locality, confirming the potential that labour market conditions experienced by remitting household members influence decisions of those remotely located in sending regions. Regardless of whether controls and fixed effects are introduced, this effect remains robust, though smaller for more elaborate specifications.

Interestingly, the white population group does not show obvious signs of wage curve behaviour with relation to labour market conditions outside their immediate locality. In fact, without controls, wages are positively related to local unemployment. However, adding fixed effects of either kind reduces this to a null or negative effect. Figure B.2 adds confidence intervals to estimates with magisterial district fixed effects: these show that the apparently negative response of white groups to distant unemployment within their own demographic also amounts to a null effect, regardless of distance. Without controls, it also appears from Figure B.1 that white wage earners do respond to black unemployment beyond their own district in the conventional manner, but fixed effects and other controls reduce this to a visible null effect, as confirmed by the confidence intervals in Figure B.2. Hence, the white population group is not responsive to information beyond their own district or demographic group. While inefficient estimates may lead to these conclusions, this is also congruent with the fact that the white population has traditionally not had spatially separated families and networks resulting from the migrant labour system. Their wage setting is also independent of labour market conditions of other groups, suggesting that they have been insulated from slack labour markets (of which other groups have borne the brunt) by past job reservation.

In all specifications, the black population group responds positively to white unemployment rates, regardless of distance. In contrast to other findings, Figure B.2 reveals that these positive estimates are statistically different from zero. This is consistent with a view that black and white labour are substitutes, as established empirically for the apartheid period (Mariotti, 2012). In cases where local

labour markets are tight for white workers (relatively low unemployment for this group), white workers are able to bargain for high wages, but this alienates the substituted black workers, who still accept low wages. During the apartheid period, job reservation was implemented to protect white workers from the competition of other black unskilled workers. However, as skills of the white population expanded relative to that of the black population, the original motivation for job reservation should by implication have disappeared. Semi-skilled white and unskilled black labour should have become complements, and job reservation would have become obsolete; the only remaining differentials in wages would have resulted from differences in educational outcomes. Mariotti (2012) does illustrate that the degree of substitutability declined substantially between these groups throughout the apartheid period, but that at no point in time did the worker types become complementary in the manufacturing sector. To some degree, this result can therefore be explained by persistent substitutability.

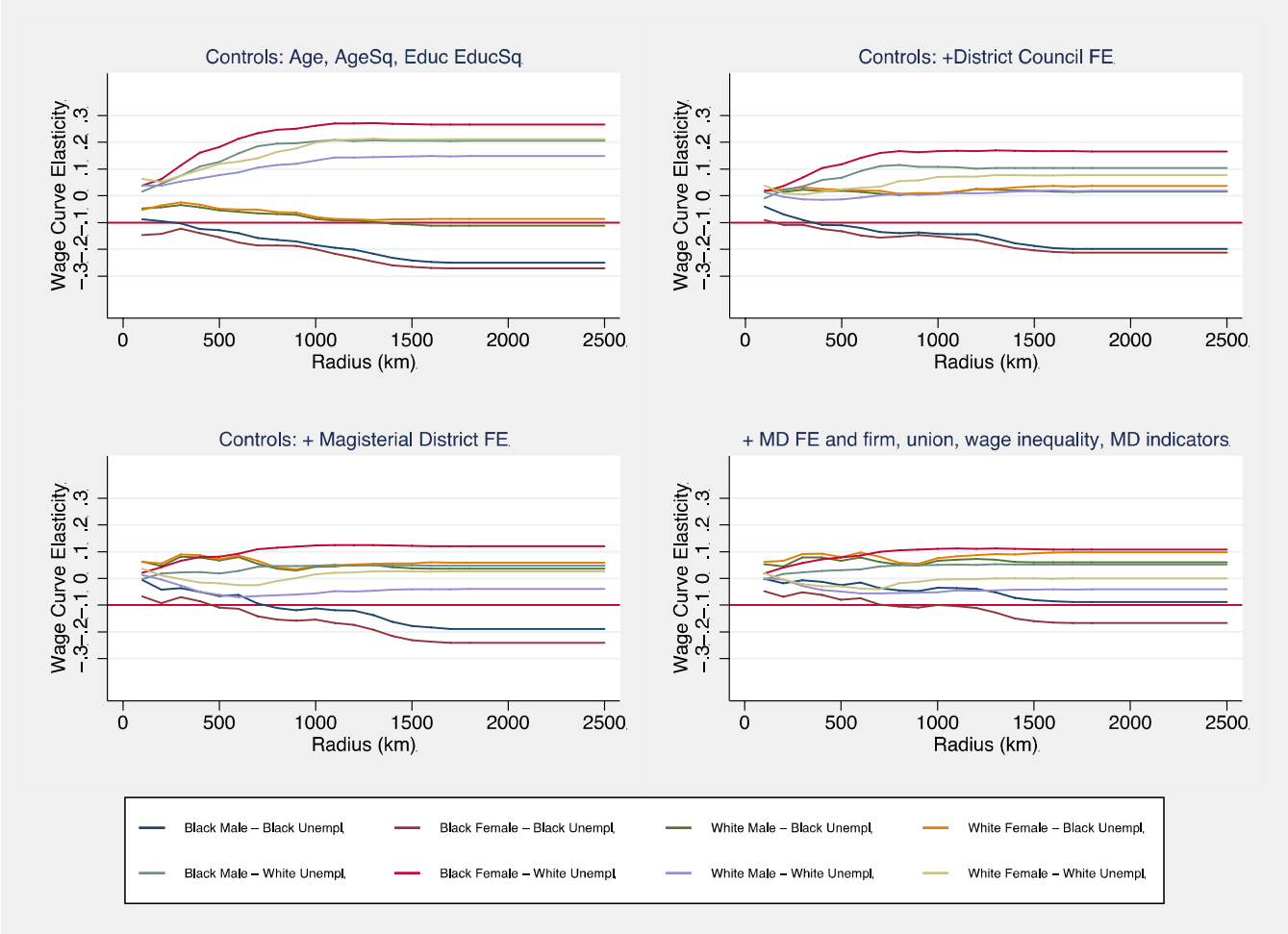

Figure B.1 Wage curve elasticities based on spatially weighted magisterial district racial unemployment rates at various radii, by demographic group

NOTES: Own calculations from LFS 2000b to 2004a. Monthly earnings measured at the individual level and deflated by a national CPI. Race-specific unemployment measured at the magisterial district level (354 in the data) according to the broad definition, and is spatially weighted. Additional controls for *magisterial district education composition, magisterial district occupation composition, magisterial district sector composition, magisterial district union density, magisterial district firm size concentration*. Standard errors are clustered at the geographic level at which unemployment is measured. Controls indicated in sub-headings. All controls from previous panel also included.

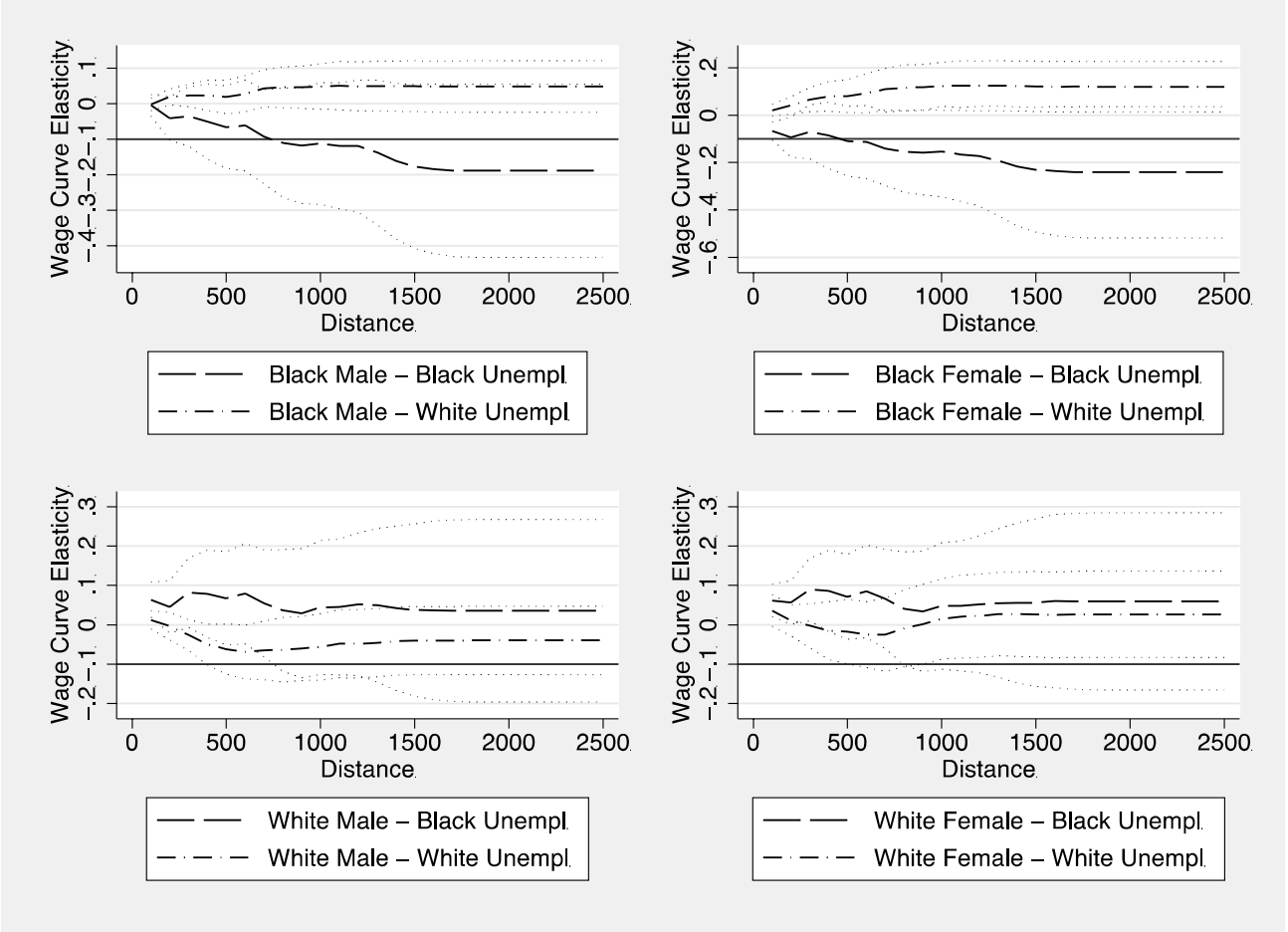

Figure B.2 Wage curve elasticities based on spatially weighted magisterial district racial unemployment rates at various radii, by demographic group, with 95% confidence intervals.

NOTES: Own calculations from LFS 2000b to 2004a. Monthly earnings measured at the individual level and deflated by a national CPI. Race-specific unemployment measured at the magisterial district level (354 in the data) according to the broad definition, and is spatially weighted. Controls for *education, age, magisterial district fixed effects, magisterial district education composition, magisterial district occupation composition, magisterial district sector composition, magisterial district union density, magisterial district firm size concentration*. Standard errors are clustered at the geographic level at which unemployment is measured.

Appendix C

Determinants of long-run wages

Chapter 2 showed that distinct short-run and long-run patterns emerged in the relationship between wages and unemployment. This appendix focuses on geographic and historic factors that are associated with long-run equilibria.

International evidence suggests that there may be differences between the long-run equilibrium and short-run adjustments in wages to high unemployment (Albaek et al., 2000). Nordic countries also experience wage inflexibility in the short-run, while the long-run spatial equilibrium is defined by regions that have high unemployment but lower relative wages.

To proceed with such an analysis, magisterial district fixed effects are extracted from the individual level regressions from specification 4 in Table 2.2. Relatively persistent variation (such as regional union density and firm size concentrations in the same period that unemployment was measured in the LFS), past institutional factors (such as homeland status), and historical shocks and initial conditions (rainfall and topography) are included to model the long-run wages[33]. Of particular importance is the notion that current settlement patterns were largely influenced by conditions in the pre-Union era. The population distribution was concentrated in similar regions to the current day, as is evident from a census map in Figure C.2 (Union of South Africa, 1912). At that time, a small thriving black agricultural community chose to cultivate crops in the regions of population concentration. They did so with some success, and some even exported their produce to the Cape Colony (Bundy, 1979). Hence, the population distribution in 1911 likely represented a spatial equilibrium that was optimal for the time. However, the Land Act of 1913 and subsequent apartheid restrictions on migration fixed these spatial patterns to a large degree, so that long-run influences likely determine the "permanent" spatial equilibrium that has resulted in South Africa. Hence, wage fixed effects (representing long-run wages) are modelled by a number of these factors, as well as the current unemployment rate, to correctly identify the long-run trade-off between wages and unemployment.

Rainfall data has been interpolated for a global terrestrial grid at half-degree intervals at a monthly frequency from 1900 to 2010 by climatologists, and is freely available for researchers (Matsuura & Willmot, 2012). These are aggregated to regional levels for South Africa, to calculate shocks in the historical and current periods. A buffer of half a degree around each regional border is created, in order to take into account that climate shocks from surrounding regions may also influence local

[33] The dependent variable varies only by district, and not across time. Similarly, many of the geographic and historical variables are time-invariant. The only time variation comes from unionization, firm concentrations and unemployment rates that are obtained from the LFS. The lack of temporal variation in long-run wages means that observations are repeated.

labour market conditions. District level rainfall shocks are calculated for the pre-1913 era, in order to proxy for potential settlement patterns (in-migration to agriculturally favourable regions) prior to formal restrictions on land ownership. It is postulated that these shocks were important in an era when agriculture dominated the economy, but that they were also made permanent by the imposition of spatial demographic segregation; potentially, therefore, the early shocks have indirect labour market impacts to this day. Recent research has shown that higher rainfall in Africa is associated with lower temperatures, which in turn promotes agricultural productivity (Fenske & Kala, 2012). Furthermore, wages in India rise in periods when rainfall increases (Rosenzweig & Udry, 2014). However, given the diminishing role of agriculture in the South African economy, these transitory mechanisms are not likely to hold in the *modern* setting. Past rainfall shocks are therefore contrasted with those during the period of analysis (2000-2004) to distinguish recent events from those which influenced long-run settlement patterns, and in turn determined long-run labour market outcomes indirectly.

Additional variables to proxy for long-run settlement patterns are also included in the analysis of the long-run equilibrium. Population densities from the 1911 census[34] are included to proxy for favourable long-run circumstances that lead to migration to particular regions (Union of South Africa, 1912). Some of the most densely populated regions of that time became apartheid homelands. This suggests that favourable initial conditions led to settlement there before any legislation by the Union government limited farmers. However, restrictions on movement prevented the initially favourable spatial distribution to continue, so that over-densification and high unemployment resulted in later years. Average terrain ruggedness for each geographic sub-region is calculated from the global raster data that Nunn & Pugo (2012) used in their cross country study of slavery. While in most countries rugged topography introduces higher transport costs and is associated with poor GDP levels, in Africa it has shielded communities and economies from historical conflicts, such as the slave trade (Nunn & Pugo, 2012). In the current context, it is postulated that pre-1913 settlement patterns were first determined by droughts, that contributed to social upheaval in the form of the *mfecane* and *difecane*[35] (Hannaford et al., 2014), which in turn prompted people to locate in geographically safe, rugged areas with high rainfall (which later became apartheid homelands).

Each of these long-run factors is related to the fixed effects from the wage curve equations, which represent long-run wages after stripping out transitory effects. The aim is to understand whether any of these variables can account for the relationship between unemployment and long-run wages. In particular, we weigh up long-run shocks against relatively newly formed institutions that govern the labour market - collective bargaining and the role of unions. Lewis (2001) highlights a lasting change

[34] Figures are weighted by the percentage of overlap between 1911 geographic regions and the magisterial district demarcations associated with the LFS. Most magisterial districts are sub-regions of 1911 demarcations
[35] While climate shocks have been a dominant narrative in the literature that establishes the roots of conflict in southern Africa, the authors are cautious of these correlations, because of spatial heterogeneity that relates to economic, cultural and political factors.

in the overall labour market, whereby black unions were deregulated in the 1970s, leading to the compression of wages ever since that period. These factors can therefore also provide explanations for the long-run spatial equilibrium.

Figure C.1 depicts the fixed effects across space. They represent the long-term, time-invariant component of wages that are not accounted for by short-run movements in the covariates. While the high long-run wage pattern for metropolitan regions (in Gauteng, Cape Town, Durban, East London, Port Elizabeth and Bloemfontein) is unsurprising, the high unemployment regions in Limpopo province also pay relatively high long-run wages. This includes many former homelands regions. It is not clear why this particular region has high wages, though some of the patterns analysed below could account for this.

Fixed effects are then regressed against local unemployment rates, industrial structure variables, as well as indicators that represent long-run shocks and settlement patterns, to establish how influential they are in describing the long-run equilibrium. Specification C1 in Table C.1 represents an unconditional long-run wage curve. The relationship between long-run wages and current local unemployment is remarkably close to -0.1, the literature gold standard. This indicates that over the long-run a spatial equilibrium emerges, whereby high unemployment regions have relatively lower wages, and vice versa. Despite short-run wage inflexibility within many local labour markets, long-run adjustments do occur, so that migration to high wage regions seem to occur if unemployment is sufficiently low to aid absorption in the receiving region. Similarly, where migration from high unemployment regions is not possible, wages appear to accommodate the mismatch. As a result, wages may present a long-run spatial clearing mechanism to respond to disproportionate labour surpluses across regions.

Specification C2 introduces historical variables that could potentially be correlated with both wages and unemployment in the long-run. However, these variables do not alter the coefficient on unemployment to a large degree, so that the relationship remains intact. Nevertheless, they raise the explanatory power of the regression substantially, indicating that historical shocks have played a role in the long-run spatial distribution and levels of wages, even if not in wage responses to high regional unemployment. Apartheid homelands experienced a 6.1 percentage point depression in long-run earnings relative to the rest of the country. This is in line with expectations, given that these regions are not only geographically isolated from economic centres, but also suffered the consequences of separate development. Results do not, however, concur with Kingdon & Knight's (2006a) finding that homelands were typically high wage and high unemployment regions, once other factors were controlled for. This will be expanded on later.

A 1 per cent increase in population density in 1911 is associated with a 0.047 per cent increase in long-term wages. Settlement patterns shortly after South Africa's unification in 1910 (but before the

1913 Land Act) therefore predated long-term wage earning capacity of regions; households settled in regions with good economic prospects from an early stage. While no attempt is made to interpret this effect causally, it does show that good conditions today were "anticipated" about a century ago, suggesting that other favourable factors were already in place at that stage.

Terrain ruggedness is associated with lower long-term wages, indicating that regions that were initially tough to settle still offer lower labour market benefits today. Hence, the transport cost or isolationist hypothesis is supported. It does not suggest that the benefits of protection from conflict that ruggedness has historically conferred on regions has manifested in long-term wages in South Africa (Nunn & Pugo, 2012). Monthly rainfall deviations (from the long-run average) in the same period that wages and unemployment were measured, contribute negatively to long-run wages, though the effect is small: a positive annual shock of 100mm of rainfall depresses wages by less than 1 per cent. This small coefficient is expected, in that short-run exogenous changes should only be incidentally correlated with long-run wages, and given that the modern labour market is less dependent on agriculture to generate high paying jobs. However, positive *historical* rainfall deviations from the long-run average (which by inference contributed to settlement decisions before the 1913 Land Act) positively influence long-run wages. Every 100mm rainfall shock per annum in the pre-Union period raises long-term wages by 12.5%. Comparing the magnitudes of current and historical rainfall shocks, the latter have a larger influence. Collectively these results suggest that settlement decisions of a century ago still manifest in current long-run labour market rewards across regions. This can be interpreted as a spatial equilibrium that was established at least a century ago, and which was then entrenched by the imposition of separate development. While these effects contribute to long-run wage levels, they do not influence the long-run wage-unemployment trade-off.

Specification C3 interacts historical population density with homeland status. Most of the coefficients are not greatly affected by this change. However, the effect of separate development becomes even clearer by these estimates. Because the coefficient on homelands now signifies a hypothetical zero population density in 1911, it represents the parts of the homelands that were not as overpopulated. They therefore displayed the capacity for wages to rise. However, the interaction effect suggests that as historical population density rose within the homelands, the long-run wage advantage is eroded. A historical map from the census (Union of South Africa, 1912) in Figure C.2 shows that the homeland regions with high long-run wages (in the North) in Figure C.1 were less densely populated in 1911 than homeland regions with low long-run wages (in the South) (Union of South Africa, 1912). Because the spatial distribution of the population has been highly persistent in South Africa, this information suggests that those parts of the homelands that experienced high initial levels of settlement also have had consistently poor labour market rewards. These low wage regions could not adjust by out-migration once homeland borders were imposed and economic conditions became

depressed. As a result, a spatial rigidity has turned *potentially* high wage regions into overpopulated areas with high levels of unemployment. However, wages have remained low.

Specification C4 turns to more recent explanations. Current union density is also associated strongly and positively with long-run wages. This does not, however, change the coefficient on unemployment, suggesting that unionisation is of short-run importance for the wage curve (as discussed in Chapter 2). District wage inequality (as measured by the variance of *log*(*earnings*)) has a negative association with long-run wages. This indicates that regions with typically high wages over the long-run also have narrower wage distributions. Taken together, these observations suggest that many districts with high long-run wages also do not have a large lower tail in the wage distribution. Some districts, therefore, specialize in high wage jobs without the proclivity to also generate low wage work.

Specification C5 introduces district-level composition of current firm sizes. Higher concentrations of larger firms raise long-run wages substantially, pointing to agglomeration regions. Interestingly, the effect of local unemployment on long-run wages diminishes somewhat to -0.085, suggesting that firm structure differences across magisterial districts are correlated with unemployment. Furthermore, this is the only determinant that also affects the long-run wage curve. Coefficients on unionization and district level inequality fall in magnitude, so that firm structure is at least partially accounting for the influence of these outcomes on long-run wages. Specification C6 confirms that with a full set of interactions, the long-run wage curve exists in districts with higher concentrations of smaller firms. If, however, only large firms were to exist within districts, no long-run trade-off between wages and unemployment would arise.

Table C.1 Long-run wage equations

Dependent variable: Magisterial district fixed effects	C1	C2	C3	C4	C5	C6
log(broad regional unemployment rate)	-0.130***	-0.133***	-0.129***	-0.128***	-0.085***	-0.153***
Homeland		-0.061***	0.210***	0.170***	0.155***	0.163***
log(1911 population density)		0.047***	0.071***	0.060***	0.038***	0.038***
Homeland x log(1911 pop density)			-0.130***	-0.107***	-0.072***	-0.073***
log(terrain ruggedness)		-0.082***	-0.063***	-0.035***	-0.031***	-0.030***
Current monthly rainfall deviations from LT avg (in 1000mm,		-0.084***	-0.062***	-0.062***	-0.042***	-0.040***
1900-1913 average annual rainfall deviations from LT avg (in		1.248***	0.882***	0.952***	0.671***	0.666***
Union density				0.818***	0.441***	0.444***
Variance of log(earnings) within region				-0.078***	-0.029**	-0.040***
Firm size concentration: 2-4					0.125***	0.143**
Firm size concentration: 5-9					0.122***	0.053
Firm size concentration: 10-19					0.371***	0.395***
Firm size concentration: 20-49					0.287***	0.433***
Firm size concentration: 50 plus					0.717***	0.859***
Firm size concentration: 2-4 x log(UR)						0.049
Firm size concentration: 5-9 x log(UR)						-0.029
Firm size concentration: 10-19 x log(UR)						0.054
Firm size concentration: 20-49 x log(UR)						0.152**
Firm size concentration: 50 plus x log(UR)						0.146***
Constant	-0.522***	-0.650***	-0.668***	-0.756***	-0.901***	-0.940***
R-squared	0.076	0.251	0.286	0.460	0.551	0.555
N	2599	2472	2472	2472	2472	2472

NOTES: * p<0.1, ** p<0.05, *** p<0.01. Own calculations from LFS Sept 2000 to LFS March 2004. Dependent variable is the set of time-invariant fixed effects estimated in Table 2.2 specification 4. Broad unemployment, current rainfall deviations, union density, variance of log(earnings) and the firm size distribution are all time-variant and measured at the magisterial district level (354 in the data). Time-invariant variables are historical in nature, and measure the district's homeland status, the district's population density in 1911, the district's terrain ruggedness and average rainfall deviations for the 1900-1913 period. Standard errors are not clustered.

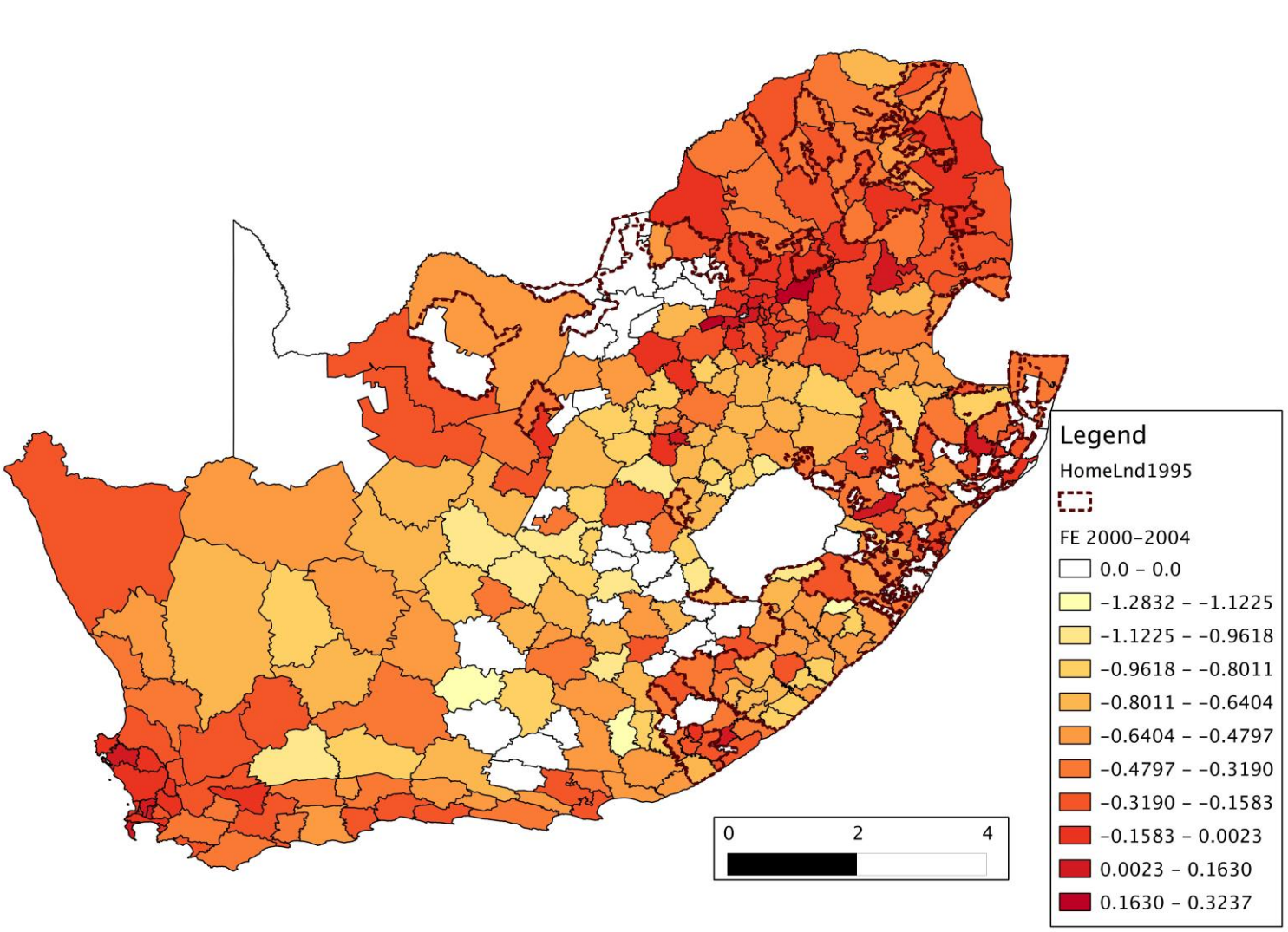

Figure C.1 Magisterial district long-run wage estimates

Source: Own calculations from LFS2000-LFS2004. Figures represent the fixed effects from Specification 4 in Table 2.2, which are measured relative to the Soweto magisterial district

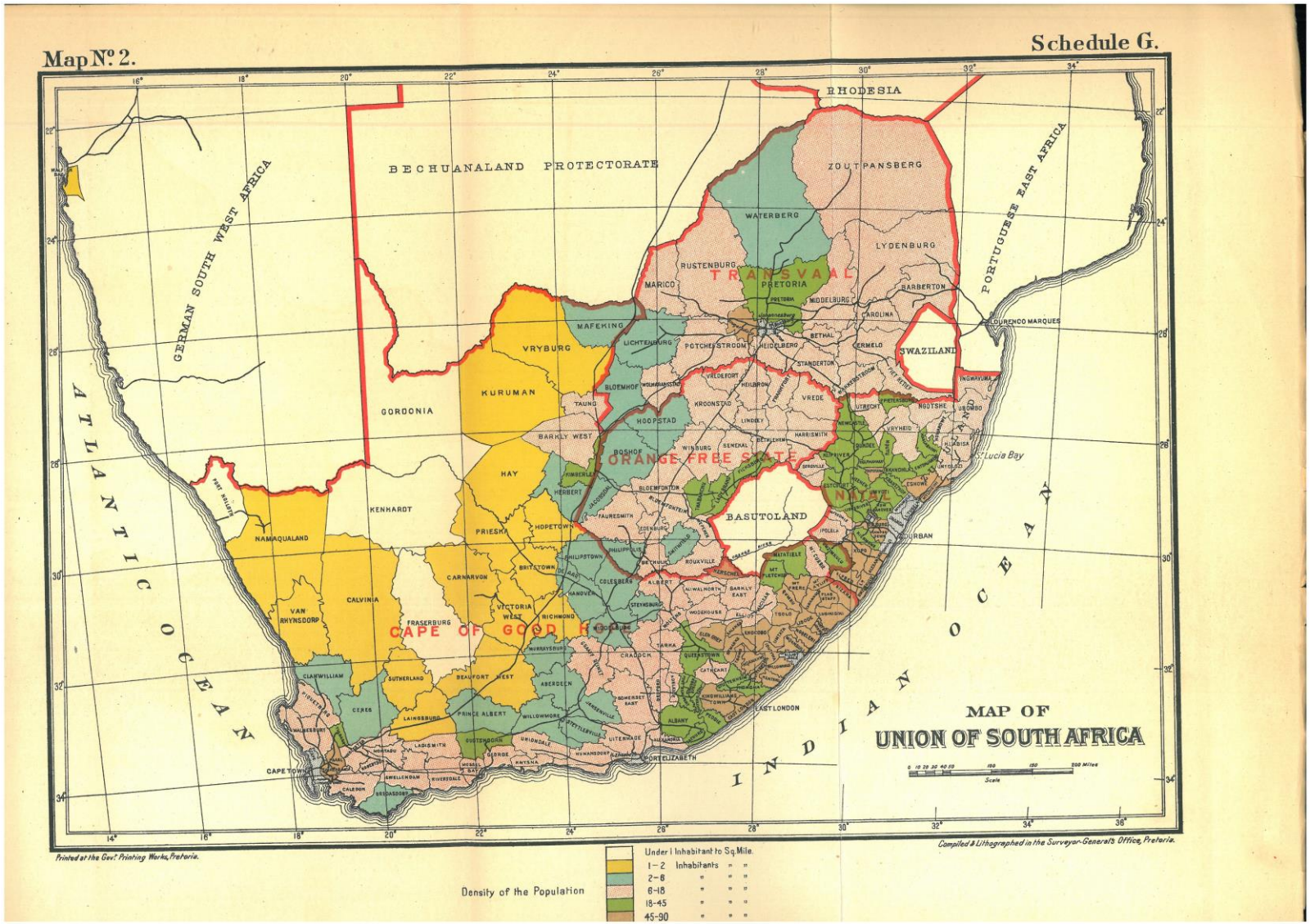

Figure C.2 Spatial distribution of population density in 1911

Source: Union of South Africa (1912)

Appendix D

Transitions in reporting childhood SES

Table D.1 Transitions in reporting childhood SES

		Childhood SES (Wave 2)						
		1	2	3	4	5	6	Total
Childhood SES (Wave 1)	1	0.382	0.350	0.200	0.048	0.016	0.003	1.000
	2	0.351	0.387	0.187	0.056	0.016	0.003	1.000
	3	0.272	0.377	0.259	0.068	0.022	0.003	1.000
	4	0.252	0.399	0.256	0.075	0.014	0.005	1.000
	5	0.257	0.347	0.297	0.069	0.020	0.010	1.000
	6	0.200	0.400	0.275	0.050	0.050	0.025	1.000
	Total	0.342	0.370	0.211	0.056	0.017	0.003	1.000

Source: Own calculations from NIDS 2008 and 2010.

Appendix E

Derivation of intrinsic estimator

This appendix provides more substance to the intuition behind the intrinsic estimator.

It is well-known that given a design matrix ($X: n\ x\ p$) and an outcome variable ($\boldsymbol{y}: n\ x\ 1$) that:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (X'X)^{-1}X'\boldsymbol{y}\ldots \quad \text{(E.1)}$$

Now, if X is not of full column rank, $X'X$ shares this property. This entails that $X'X$ is singular, and hence the OLS solution is not uniquely determined. Suppose that $X: n\ x\ p$ is of rank $k < p$, where $k = A + P + C - 3$ and $p = A + P + C + 1$, and A, P and C are the number of ages, periods and cohorts in the data. The singular value decomposition is able to uniquely decompose X as in equation E.2 (Johnson & Wichern, 2002).

$$X = UDV' = \sum_{i=1}^{A+P+C+1} \lambda_i^{0.5}\, \boldsymbol{u}_i \boldsymbol{v'}_i = \sum_{i=1}^{k} \lambda_i^{0.5}\, \boldsymbol{u}_i \boldsymbol{v'}_i + \sum_{i=k+1}^{A+P+C+1} 0 * \boldsymbol{u}_i \boldsymbol{v'}_i = \widetilde{U}\widetilde{D}\widetilde{V}' + \mathbf{0} \ldots \text{(E.2)}$$

where $\lambda_i^{0.5}$ are the singular values of X or equivalently λ_i are the eigenvalues of both $X'X$ and XX'. The first k eigenvalues are non-zero, and the last $p - k$ are zero, resulting in the singularity of X; $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the column vectors of $U: n\ x\ n$ and $V: p\ x\ p$ respectively. U and V are orthogonal matrices containing eigenvectors of XX' and $X'X$ respectively, while D is an $n\ x\ p$ matrix with $\lambda_i^{0.5}$ comprising the first k "diagonal" elements ($[i, i] = [1,1], [2,2] \ldots [k, k]$), and all other elements equal to zero. The first k column vectors of U and V correspond to the non-zero singular values.

Due to the $p-k$ singular values which equal zero (these are the root of singularity of the matrix), X can be expressed in terms of only k columns of U and V respectively, rather than p, without loss of information. This is denoted by the multiplication of the reduced matrices $\widetilde{U}\widetilde{D}\widetilde{V}'$. Each of these is a sub-matrix, which include only the first k columns of the original matrices, and therefore exclude the "redundant" information which result in the singularity. This is analogous to principal components analysis, where the eigenvectors of a covariance matrix (which correspond to non-zero eigenvalues) summarize the data adequately without loss of information. It must be noted, however, that X's column space continues to have dimension p while the rank of X remains $k < p$. Hence, even with the use of the reduced set of bases, X is still singular. Using a transformation, however, solves this problem.

Consider the transformation of X, which follows because $\tilde{V}$ is orthogonal:

$$\tilde{X} = X\tilde{V} = \widetilde{U}\widetilde{D}\widetilde{V}'\tilde{V} = \widetilde{U}\widetilde{D} \text{ ... (E.3)}$$

Now the transformed X is of rank and dimension k, and thus excludes the null space of X, lending it to the construction of a linear estimator. Consequently the transformed X is non-singular, and it is possible to compute a preliminary estimator of the coefficient vector:

$$\widehat{\boldsymbol{\beta}}_{SVD} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\boldsymbol{y}=\left(\tilde{V}'X'X\tilde{V}\right)^{-1}\tilde{V}'X'\boldsymbol{y}=\left(\widetilde{D}'\widetilde{U}'\widetilde{U}\widetilde{D}\right)^{-1}\widetilde{D}'\widetilde{U}'\boldsymbol{y}=\left(\widetilde{D}'\widetilde{D}\right)^{-1}\widetilde{D}'\widetilde{U}'\boldsymbol{y}$$

$$\Rightarrow \widehat{\boldsymbol{\beta}}_{SVD} = \widetilde{D}^{-1}\widetilde{U}'\boldsymbol{y} \text{ ... (E.4)}$$

because $\widetilde{U}$ is orthogonal and $\widetilde{D}$ is diagonal.

This estimator, however, is not comparable to the OLS estimator, as it is estimated within a rotated vector space, and coefficients will not have the usual interpretation. A rotation back to the original space is therefore required. To recover the coefficients with respect to the original vector space, it is necessary to note that:

$$\widetilde{D}\widehat{\boldsymbol{\beta}}_{SVD} = \widetilde{D}\widetilde{D}^{-1}\widetilde{U}'\boldsymbol{y} = \widetilde{U}'\boldsymbol{y} \text{ ... (E.5)}$$

Now, when X is of full rank, using E.3 and E.5, OLS estimates can be re-written as:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (X'X)^{-1}X'\boldsymbol{y}=\left(\tilde{V}\widetilde{D}\widetilde{U}'\widetilde{U}\widetilde{D}\widetilde{V}'\right)^{-1}\tilde{V}\widetilde{D}\widetilde{U}'\boldsymbol{y}=\left(\tilde{V}\widetilde{D}\widetilde{U}'\widetilde{U}\widetilde{D}\widetilde{V}'\right)^{-1}\tilde{V}\widetilde{D}\widetilde{D}\widehat{\boldsymbol{\beta}}_{SVD}$$

$$= (\tilde{V}\widetilde{D}\widetilde{D}\widetilde{V}')^{-1}\tilde{V}\widetilde{D}\widetilde{D}\tilde{V}'\tilde{V}\widehat{\boldsymbol{\beta}}_{SVD} = \tilde{V}\widehat{\boldsymbol{\beta}}_{SVD} \text{ ... (E.6)}$$

When X is of full rank, OLS provides a unique estimator, which can be expressed in terms of the SVD estimator. Since there is no null space in this case, the SVD estimator does not project into another space whatsoever. When X is not of full column rank, $\widehat{\boldsymbol{\beta}}_{OLS}$ does not have a unique solution due to the

singularity of $X'X$, and cannot be computed in the standard manner. However, the SVD estimator can be calculated and provides one such solution, based on only the column space of X, and excluding its null-space. Using only the information contained in $\tilde{V}\widehat{\boldsymbol{\beta}}_{SVD}$ provides an estimator for the coefficient vector, though, as discussed in Chapter 4, it is not unique: other solutions would add some weight to the null space.

Using the first k columns of V, it is therefore evident that $\widehat{\boldsymbol{\beta}}_{SVD}$ can be rotated into the original vector space, and it is possible to obtain an estimator comparable in interpretation to OLS. This vector now includes a parameter estimate for the variable, which resulted in the singularity of X. This procedure entails that no pre-meditated behavioural restrictions on variables are required to artificially force full column rank (this includes variables which exhibit perfect multicollinearity with other variables, but also the "reference" categories in a dummy variable regression, which are usually assumed to equal zero).

Hence, the intrinsic estimator can be calculated as:

$$\widehat{\boldsymbol{\beta}}_{IE} = \tilde{V}\widehat{\boldsymbol{\beta}}_{SVD} = \tilde{V}\tilde{D}^{-1}\tilde{U}'\boldsymbol{y} \quad \text{... (E.7)}$$

The intrinsic estimator would be equivalent to OLS if there had been no collinearity in the design matrix X. Hence, the intrinsic estimator can be viewed as a special case of OLS, but which can also be estimated without the otherwise necessary Gauss-Markov assumption of perfect collinearity.

Appendix F
Distributions of simulated APC estimates

The appendix presents more in depth results from the simulation study using the interactive data generating process. Output for each of the additive models is presented, with a box plot for each of the age, period and cohort coefficients that are estimated based on the 1000 simulations. In each case, the top left panel repeats the simulated age-cohort structure. The top right panel gives the distribution of age coefficients/predictions. The bottom left panel does the same for cohort estimates, while the bottom right panel repeats this for period estimates.

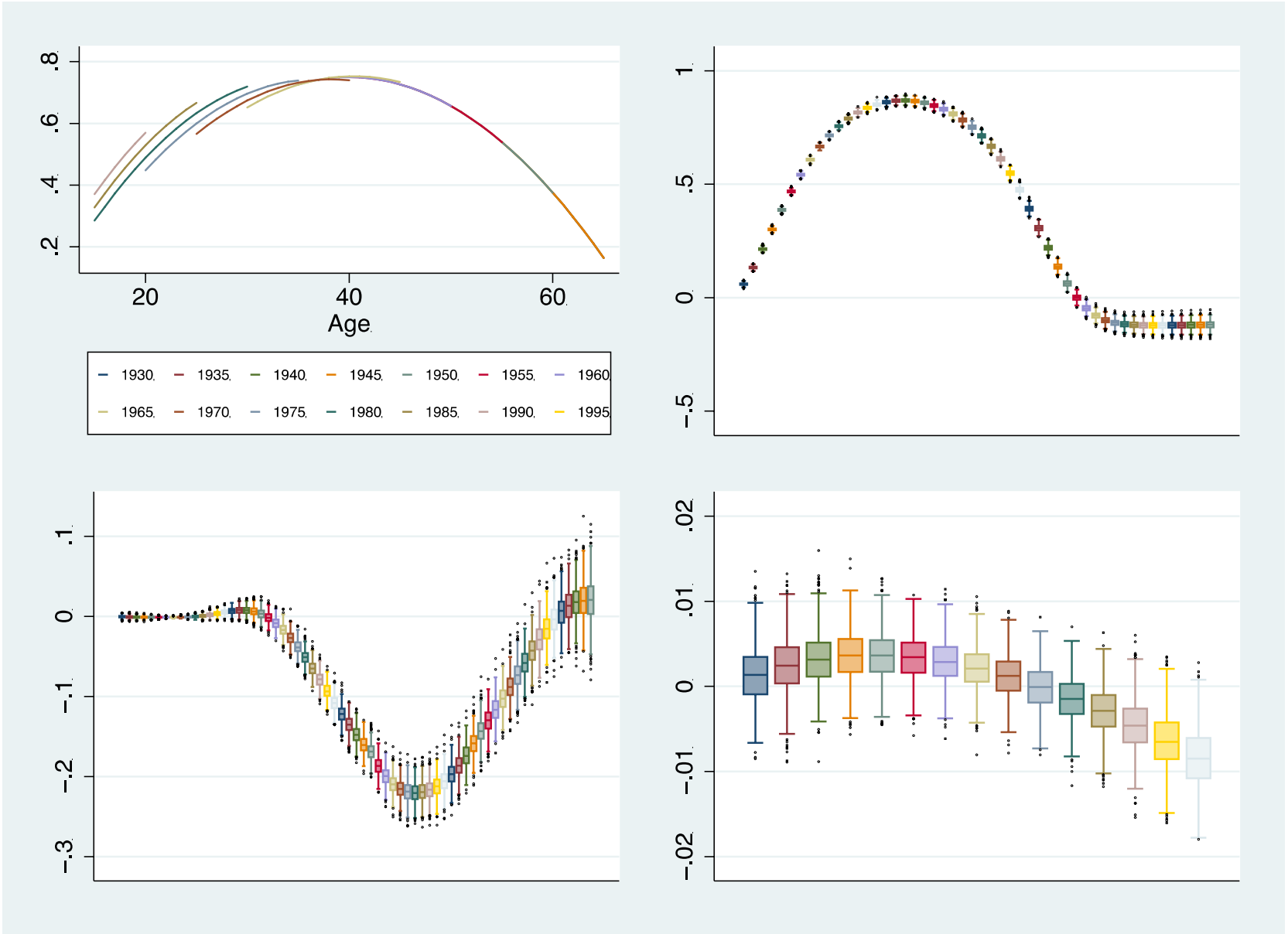

Figure F.1 Distributions of coefficient estimates using Deaton-Paxon restriction on interactive simulated data

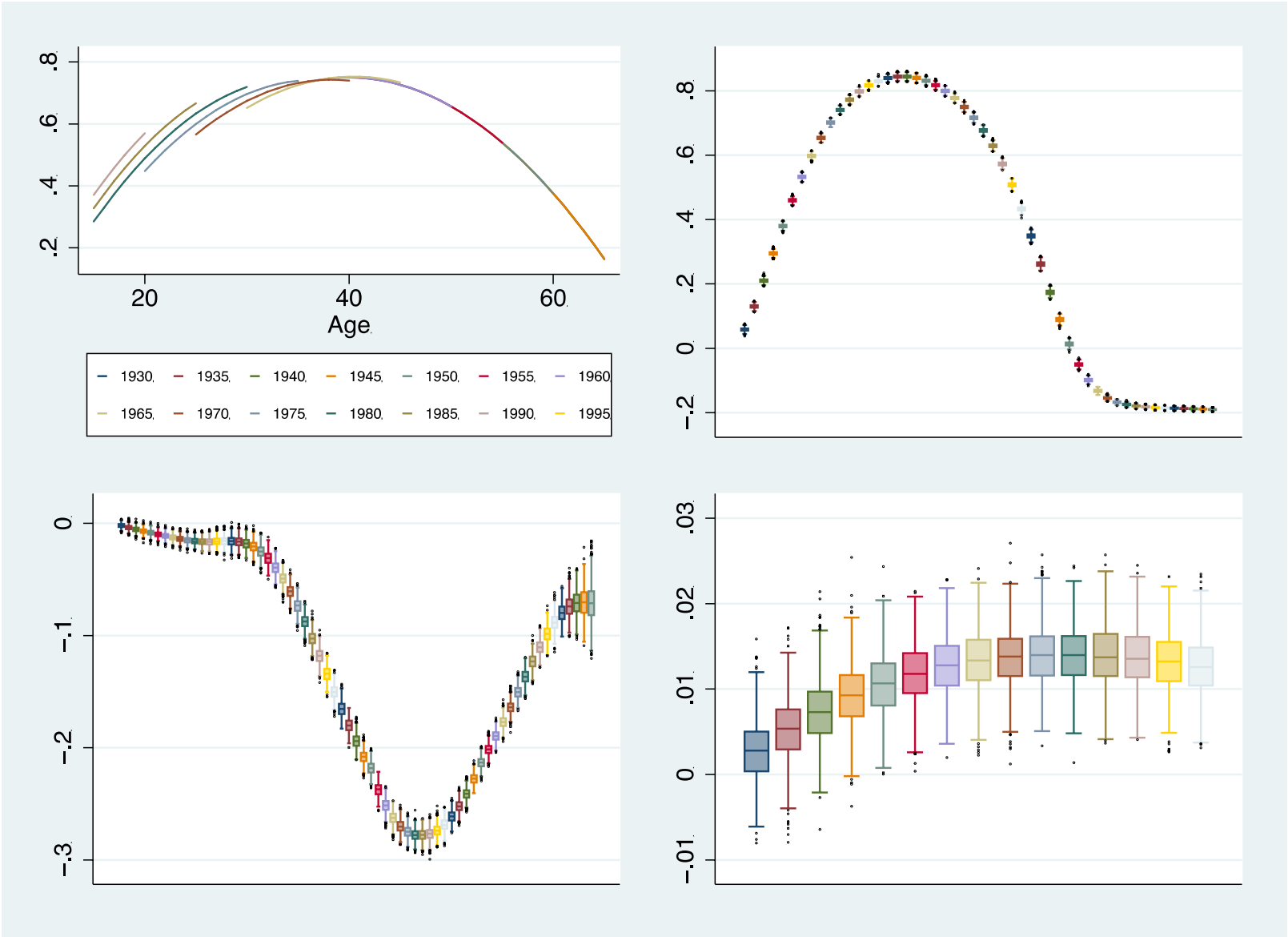

Figure F.2 Distributions of coefficient estimates using maximum entropy estimator on interactive simulated data

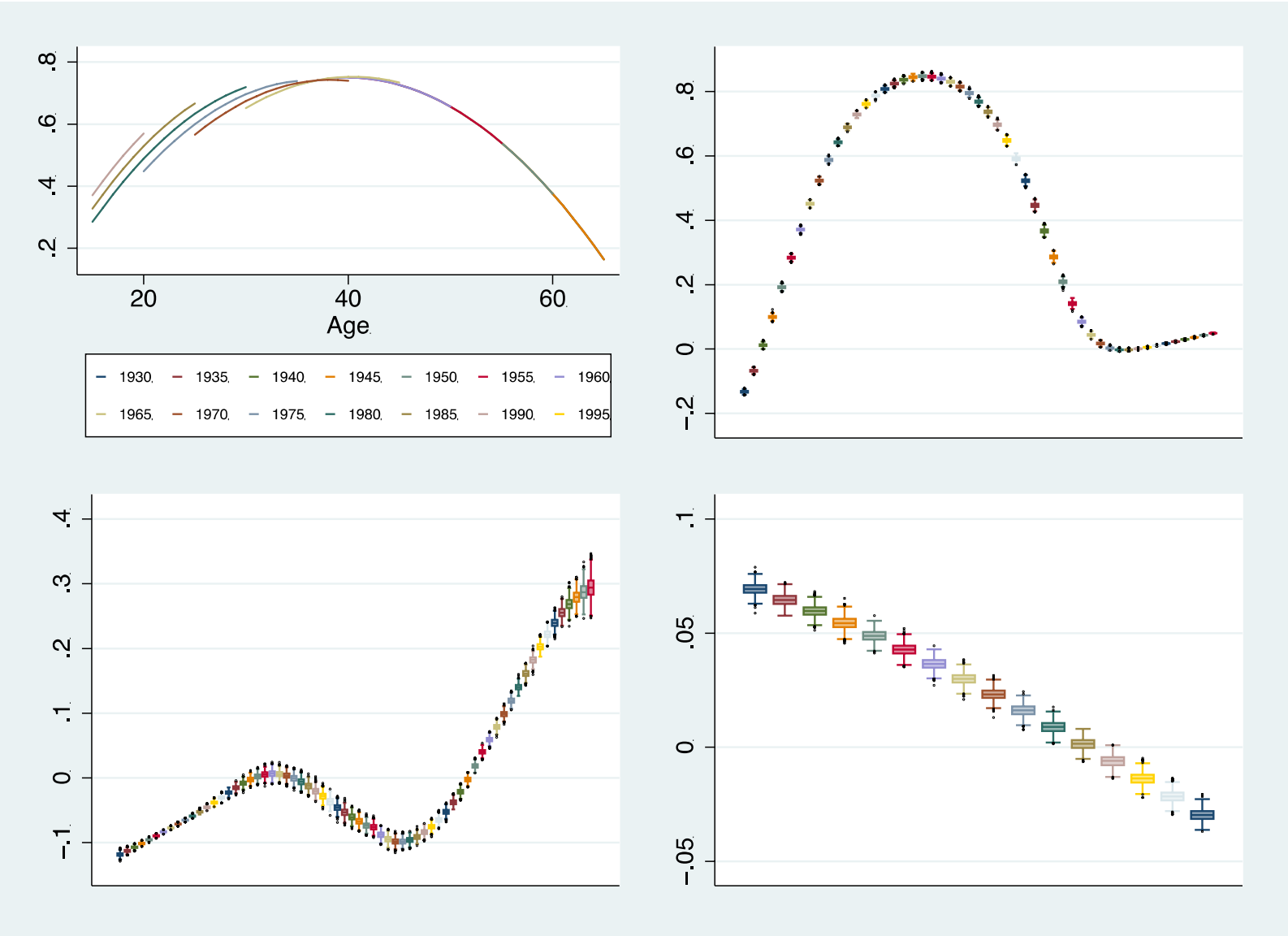

Figure F.3 Distributions of coefficient estimates using intrinsic estimator on interactive simulated data

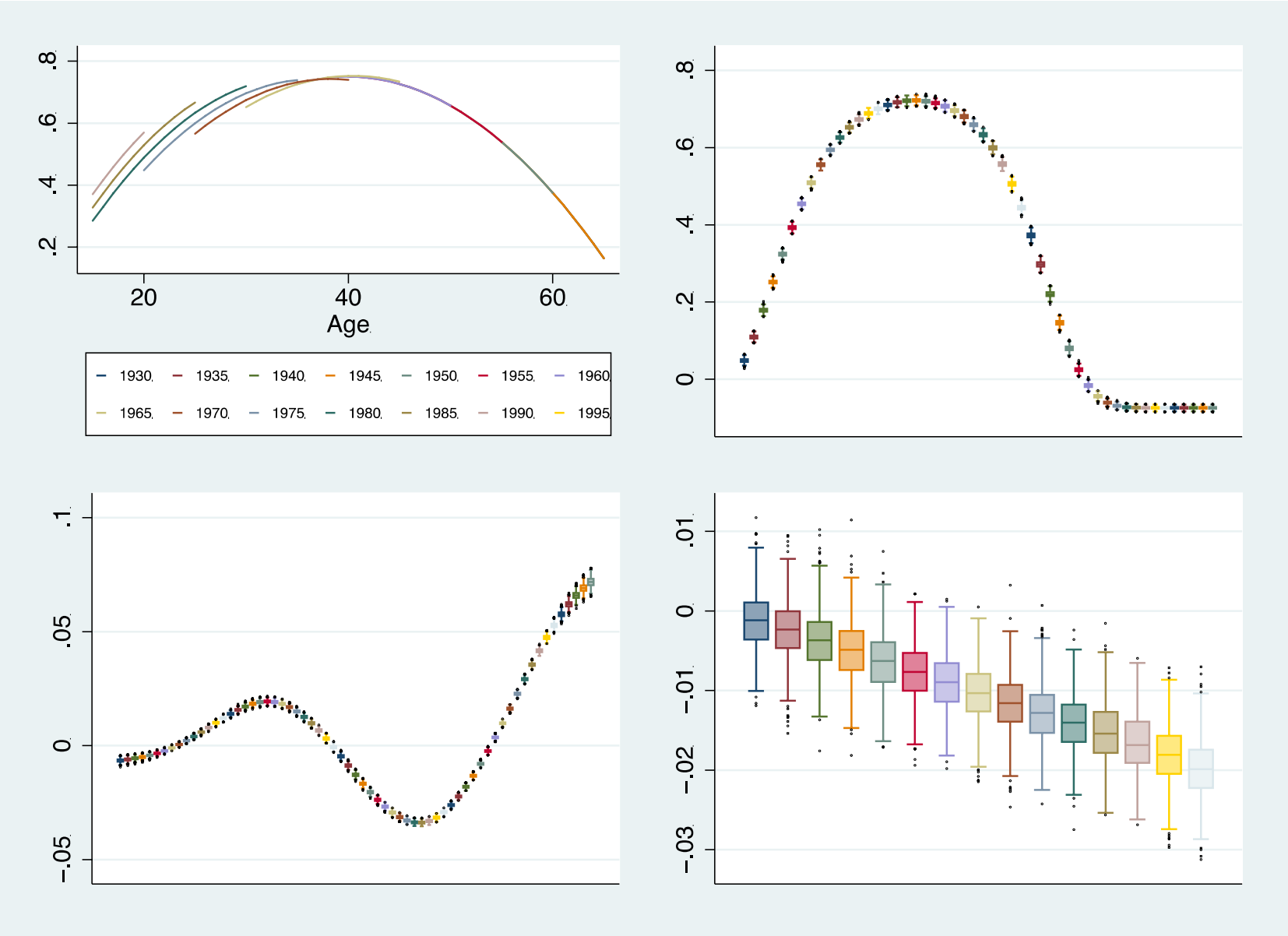

Figure F.4 Distributions of coefficient estimates using semi-parametric estimator on interactive simulated data

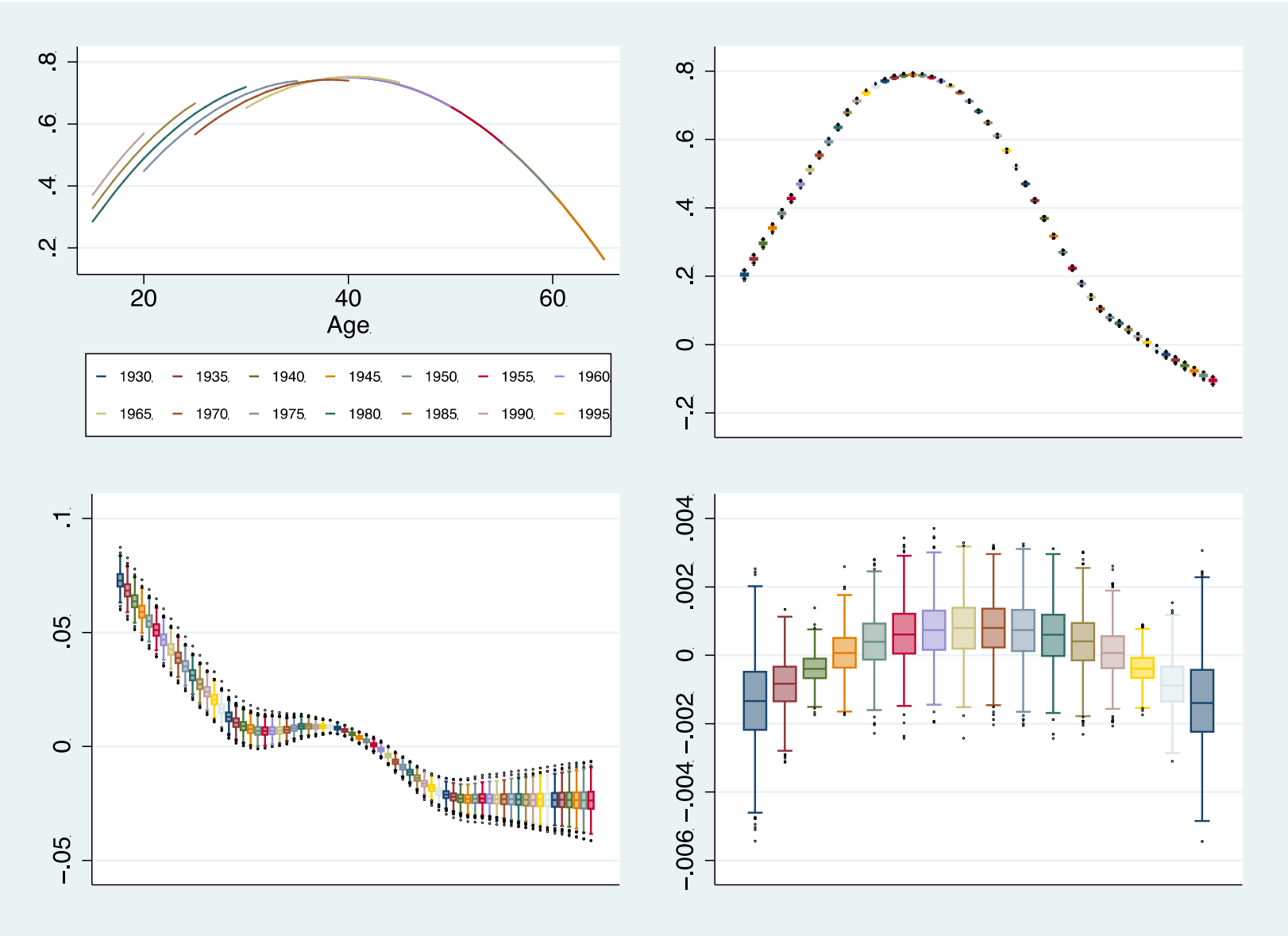

Figure F.5 Distributions of coefficient estimates using additive Generalized Additive Model on interactive simulated data