Simulation analysis of ageing and inequality

[31.8.2011]

Paul Lambert, Meik Poschen and Guy Warner

NeISS, www.neiss.org.uk
(National e-Infrastructure for Social Simulation)

A 'Fringe workshop' to the 'Social Stratification Research Seminar', Univeristy of Stirling, 31st August – 2nd September 2011

Workshop: Simulation analysis of ageing and inequality (4-6pm, 31st August 2011, LT2V1)

Presentation from Paul Lambert:

1) NeISS – resources for social simulation and the 'Ageing and Inequality' exemplar (www.neiss.org.uk)
2) Social simulation and data management
3) Specification and early results from the 'Ageing and Inequality' microsimulation

Presentation from Meik Poschen:

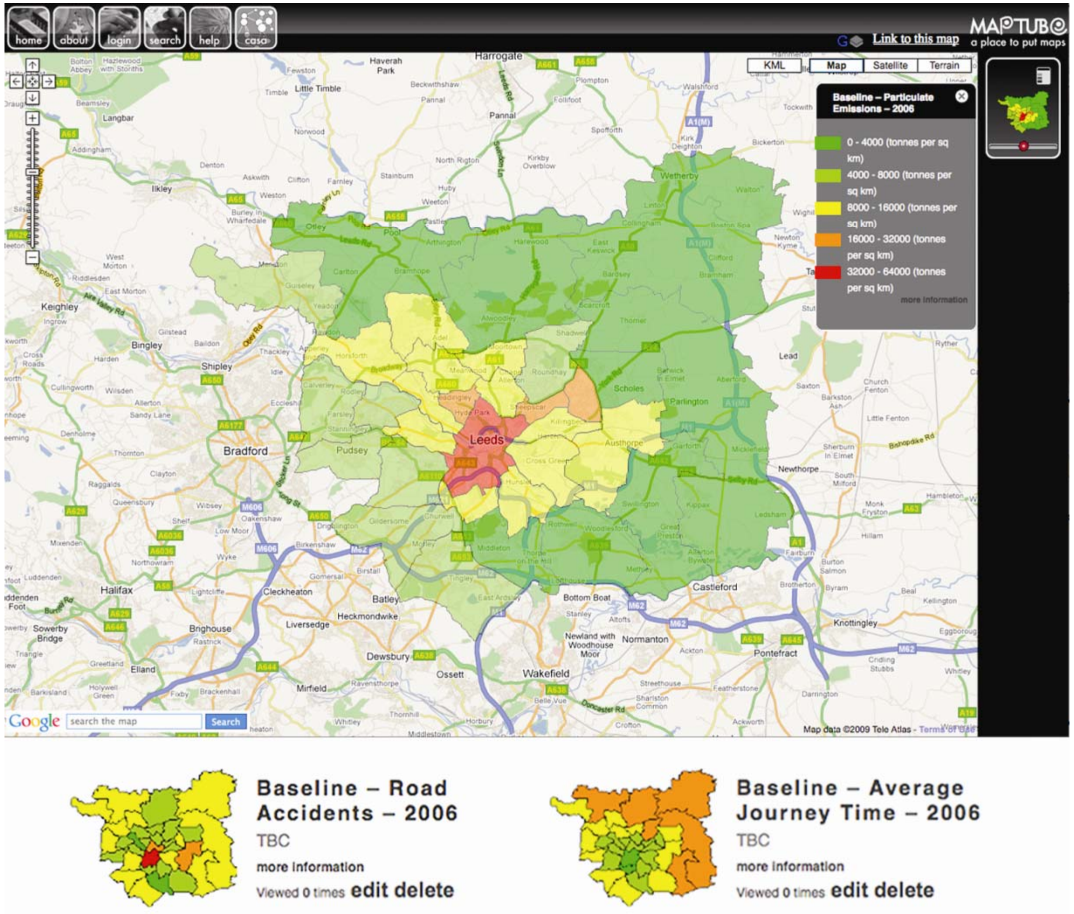
1) NeISS: A look at the live simulation specification tool

Discussion/User communities (MP and PL)

1) www.neiss.org.uk

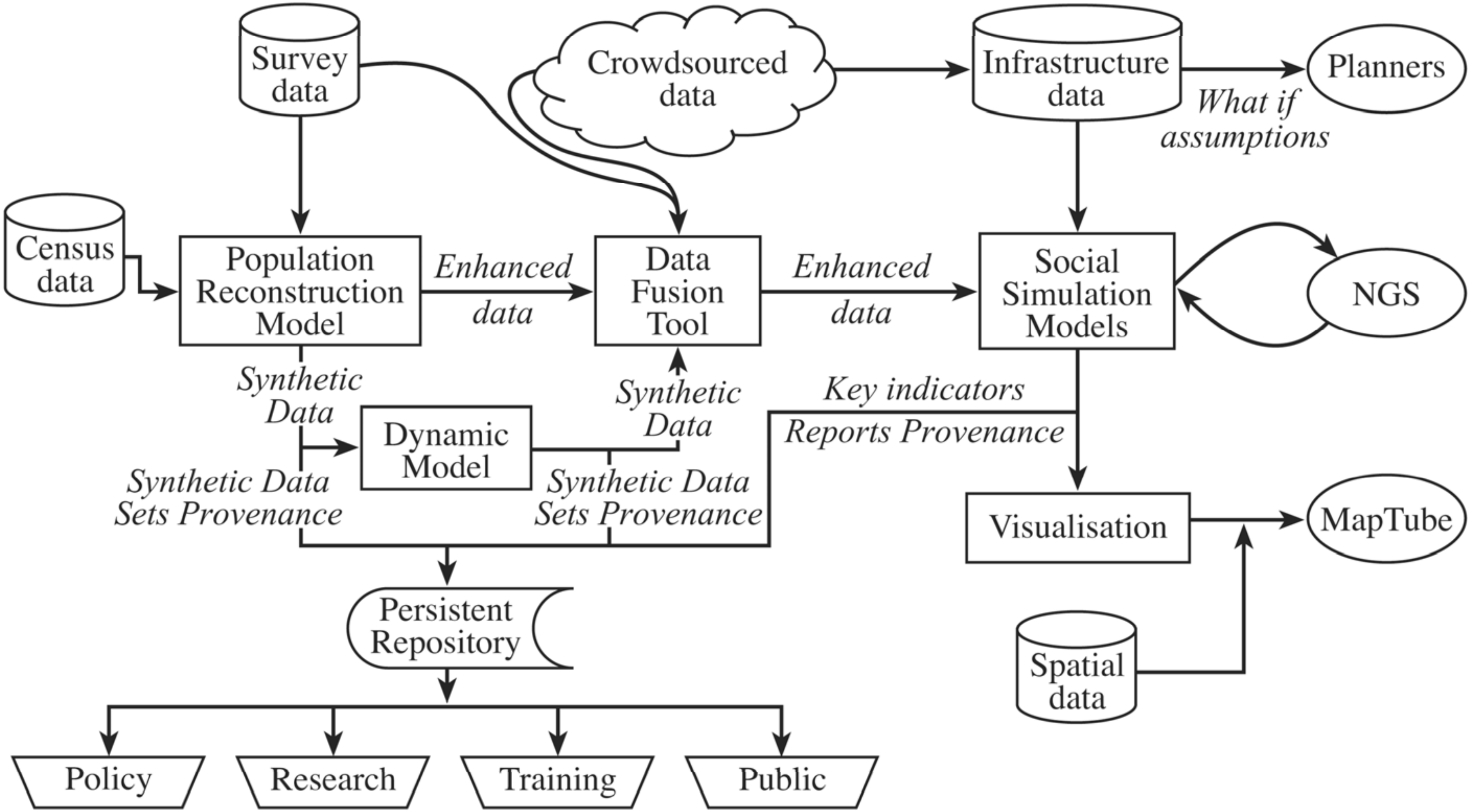
A JISC initiative (2009-12) on collaborative research infrastructure in the UK

NeISS



- ➢ National e-Infrastructure for Social Simulation
- Expert led simulation demonstrations
- Combining data resources
- 5 exemplar application areas
- Workflows for the simulation analysis (modify and re-specify existing simulation templates)

See Birkin et al. (2010) (includes image source)



Birkin et al. (2010: 3808)

Contributions of the ‘NeISS’ project

- Accessing live / newly updated socio-economic/demographic/geographic data
- Running/supporting complex simulation models with high computational requirements
- Summarising complex outputs (e.g. geographical mapping)
- Allowing flexible data management (e.g. define/compare alternative measures of occupational position, education)
- Allowing multiple specifications of related models for comparisons (e.g. vary a few parameters and re-run)

➢ ***…application areas concerning demographic projections; socio-economic inequalities; geographical influences; model visualisations…***

Application on 'Ageing and inequality'

- Sociological and econometric research agendas studying the circumstances of social inequality

  who is advantaged/disadvantaged; why is that?

- We increasingly acknowledge the potential influences of demographic/socio-economic transformations

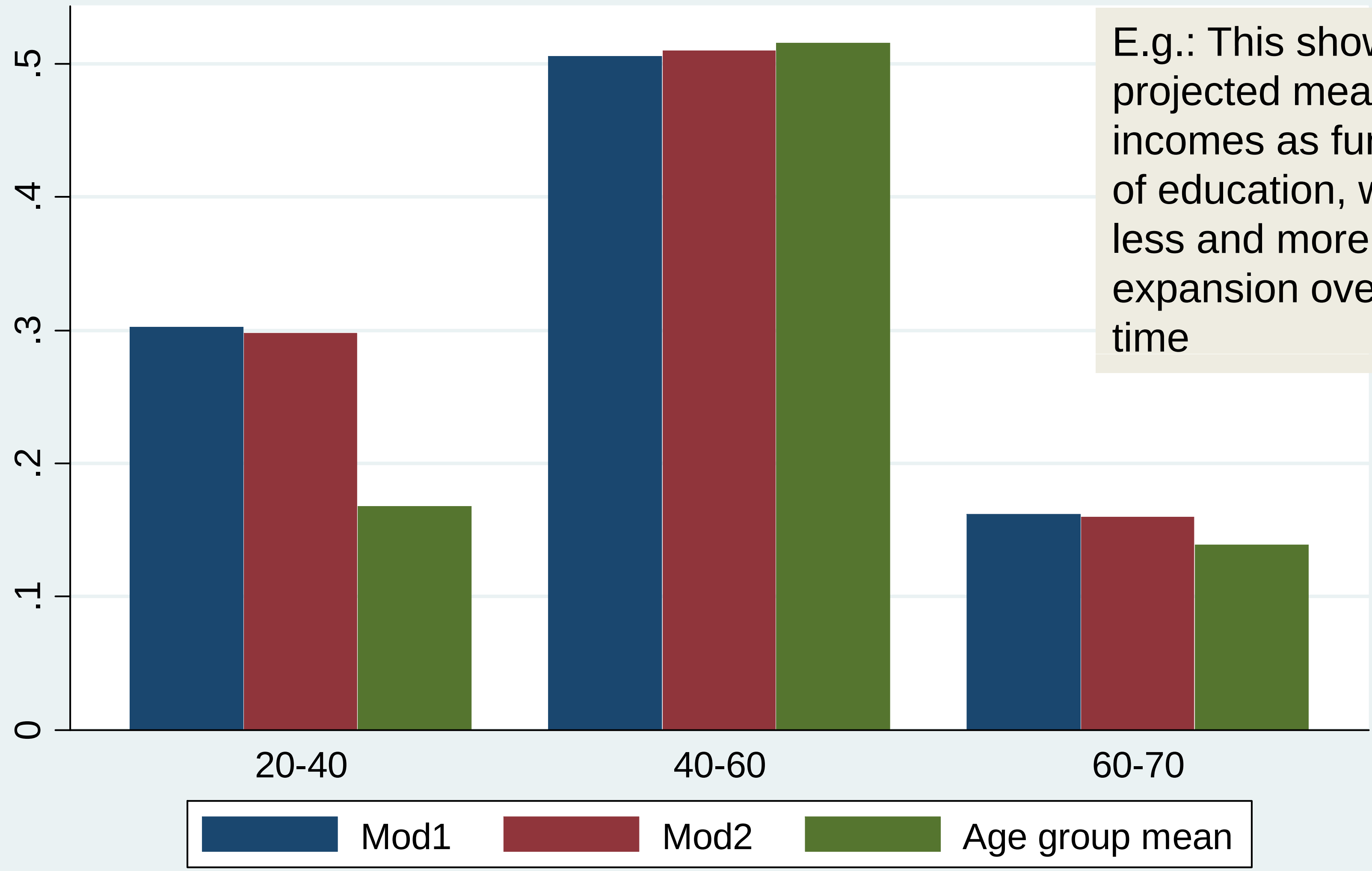
  ageing population; changing family structures;

  educational expansion; immigrant influxes

- Ample long-running longitudinal survey data resources in UK

  e.g. BHPS; GHS; LFS; 'Slow Degrees' dataset

- Many previous simulation analyses compare the effects of social changes on social inequalities (e.g. Zaida et al. 2009),
  - to our knowledge, there is little attention paid to, or sensitivity analysis of, alternative measures of social structure and inequality other than income - such as of occupations, educational levels



E.g.: This shows projected mean incomes as function of education, with less and more uni. expansion over time

BHPS w17, Males in employment aged 20-70
Model 1: Income=0.39Educ + 0.1Age -0.001Age2 + 0.13Age*Educ
Model 2: Income=0.17Educ + 0.1Age -0.001Age2 + 0.26Age*Educ

Some possible research questions

- How age-qualifications links impact trends in social inequality
  - Mass education; admissions policies; cognitive versus sheepskin effects
- How will (high/low qualified) cohort-specific immigrant influxes impact upon regional age-occupation-qualification distributions
  - Simulation: increase or decrease proportions within birth cohorts/ethnic groups/regions/sectors with certain qualifications
- How will fine-grained industrial sector transformations impact different age cohorts and subsequent stratification positions (e.g. rise of the 'cultural industries')
  - Simulation: Modify national and/or local industrial distributions and project forward over time
- How is long term wealth accumulation influenced by longer life expectancies (e.g. changing inheritance patterns; longer pension dependence)
  - Simulation: Model and modify income through work and through inheritance as it influences relative social position at a national level

2) Social Simulation and Data Management

- Simulation analysis intrinsically involves:
  - *data simulated (i.e. constructed)* forward in time
  - *analysis* to summarize progression through time
- Microsimulation
  - Using starting parameters from existing micro-data to model predicted values in the future
    - Year on year values carried through, using transition probabilities
    - E.g. Gilbert & Troitzsch (2005); Gilbert (2008); Zaidi et al. (2009)

…The starting parameters (and projected data) are all subject to socio-economic measures, which are negotiable…

The contribution of simulation

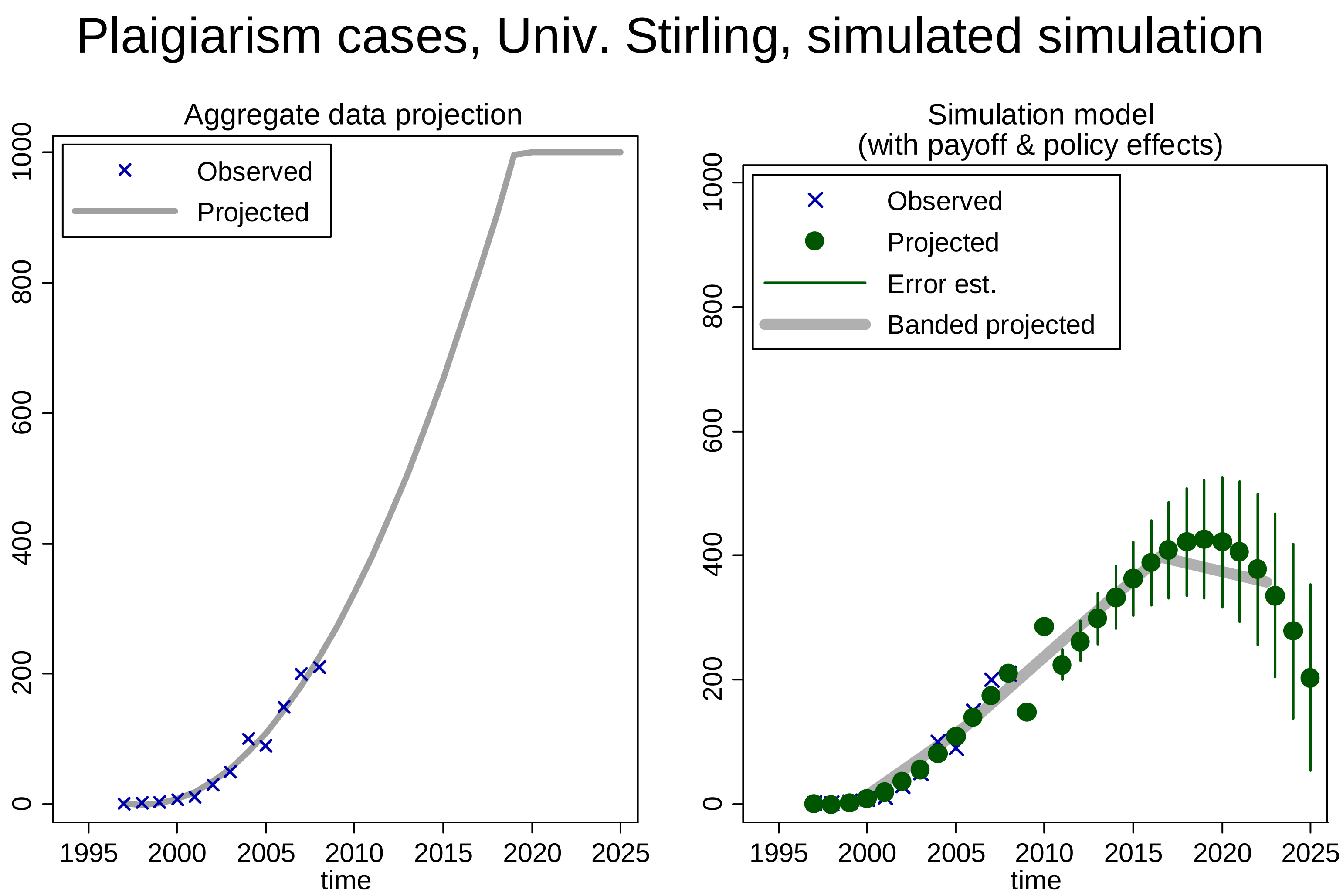
- The general contribution is to model forwards in order to see plausible patterns within complex/responsive systems
  – Needs a good model of influences, projected influences, contextual effects

    (serious models take a lot of work – e.g. Euromod; SAGE)

  – We ordinarily try various inputs to the system

    (e.g. what would happen if we did X)

The challenges of simulation…



Panel 1, pred = cons + t1*time + t2*time2.
Panel 2, pred = (cons + t1*time + t2*time2)*((2000-(cons + t1*time + t2*time2))/1900) - (100*(pred[_n-1] >= 200))

‘Data management’ choices

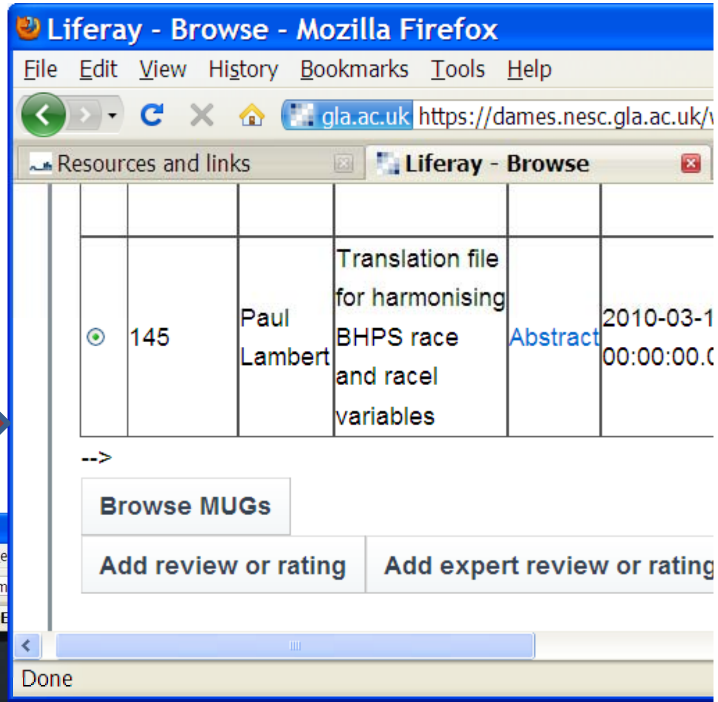
- In the social sciences, most concepts can potentially be measured by several different indicators
  - *E.g. occupation-based measures; education; family type; individual/household controls*
- In ‘e-Social Science’ there are several initiatives in facilitating operationalisation of measures
  - ***Access to data****: e.g. GEODE/GEEDE/GEMDE; Methodbox*
  - ***Tools for linking data****: DAMES; Taverna; MyExperiment*

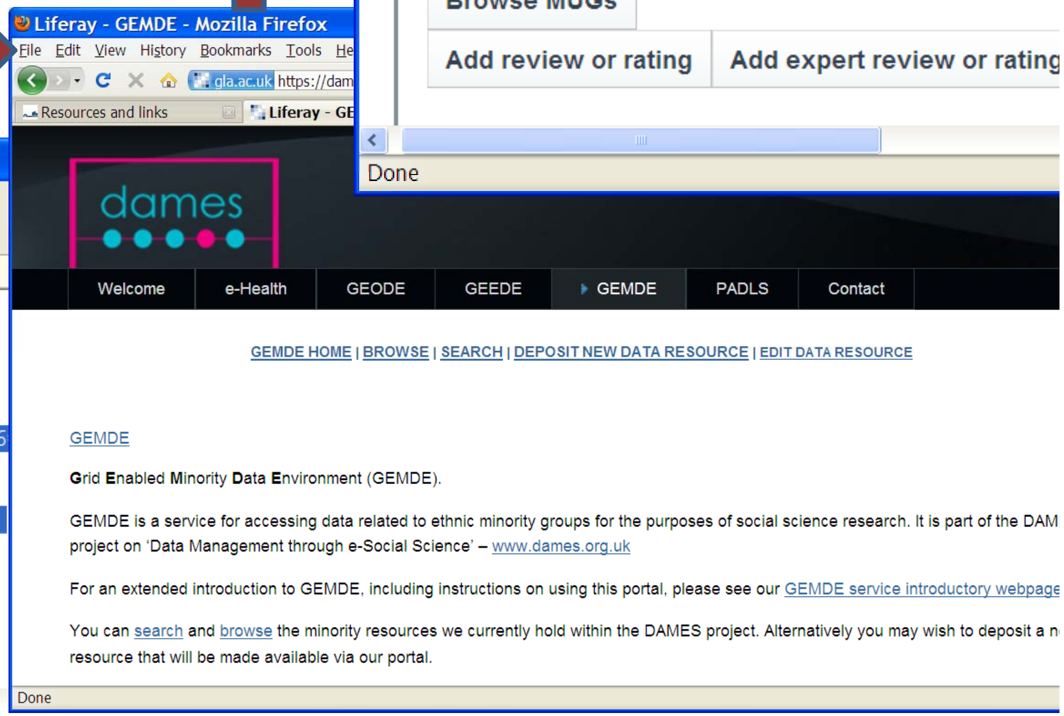
GESDE: online services for data access/ coordination/organisation

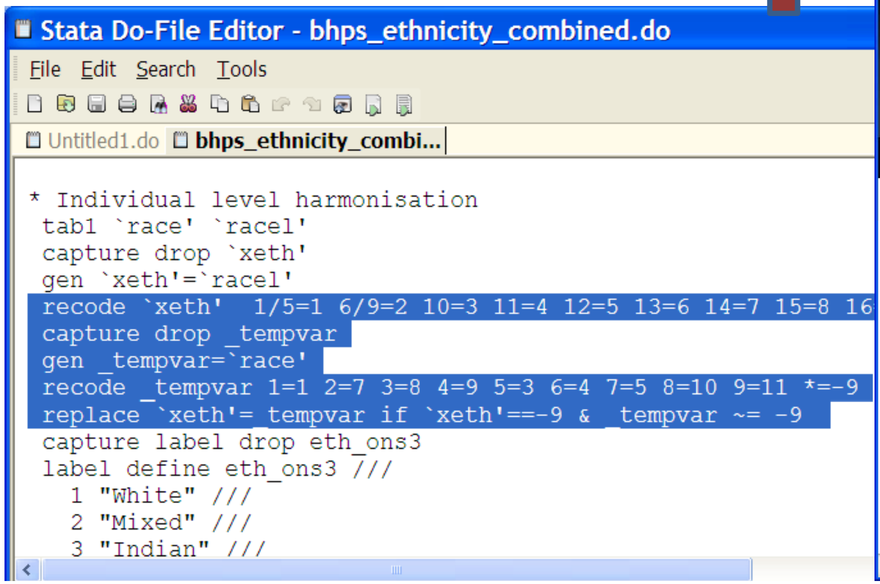
www.geode.stir.ac.uk / www.dames.org.uk/gemde / www.dames.org.uk/geede

Specialist information on occupations, educational qualifications, ethnicity

Data on recodes, standards, & harmonisation







DAMES – data fusion tool

```
universe=vanilla
executable    = /usr/bin/R
arguments     = --slave --vanilla --file=match_merge1.R --args
/home/pl3/condor/condor_5/wave1.dta
/home/pl3/condor/condor_5/wave17.dta
/home/pl3/condor/condor_5/bhps_combined.dat pid wave file pid
wave file
notification = Never
log           = test1.log
output        = test1.out
error         = test1.err
queue
```

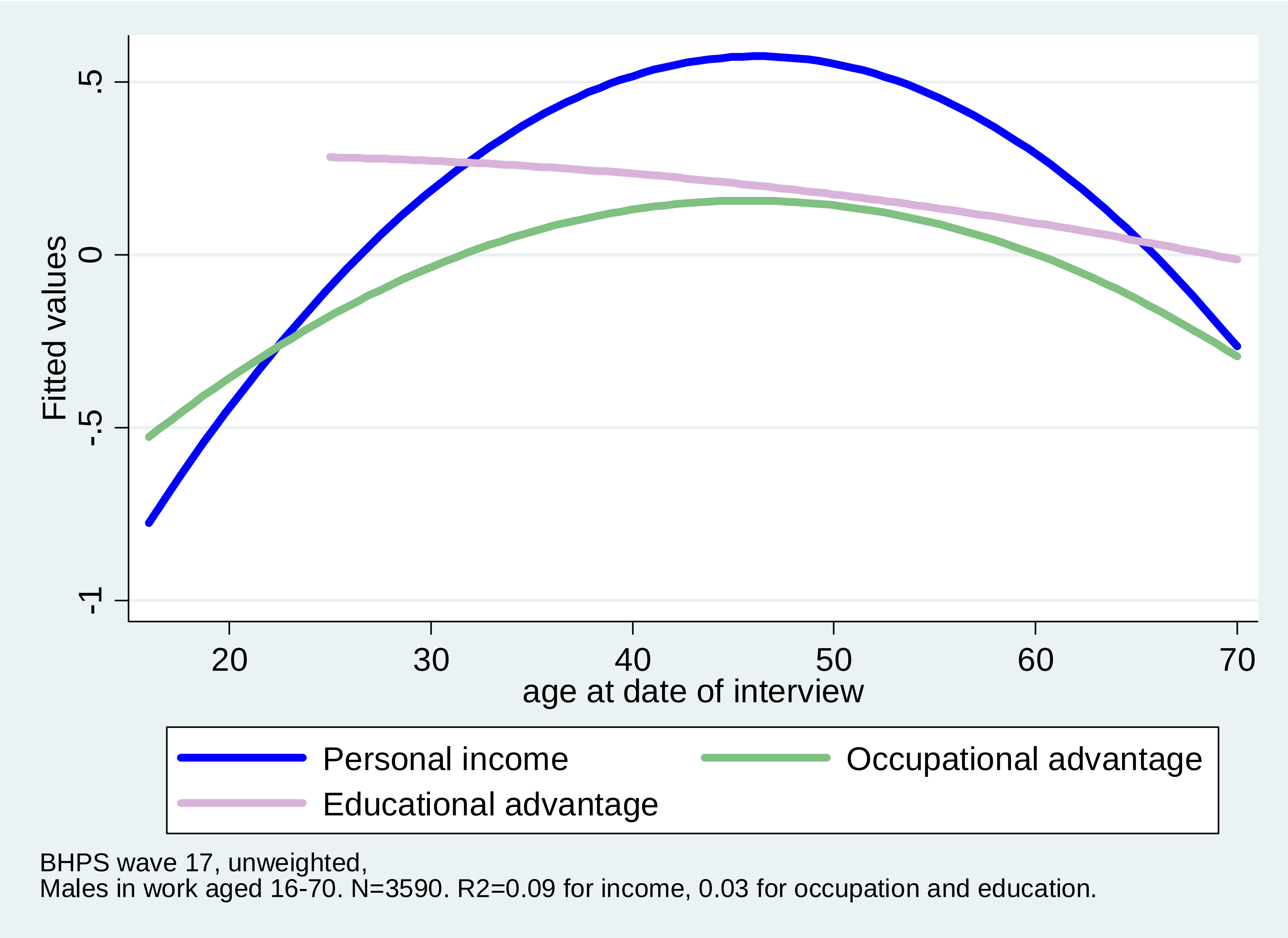
- Irods filestore system
- Condor job submission tool to run pre-defined file linkages or other tasks

```
args <- as.factor(commandArgs(trailingOnly = TRUE));
options(useFancyQuotes=TRUE)
fileAinp <- as.character(args[1])
fileBinp <- as.character(args[2])
fileCout <- as.character(args[3])
##
library(foreign)
fileA <- read.dta(fileAinp, convert.factors=F)
fileB <- read.dta(fileBinp, convert.factors=F)
nargs <- sum(!is.na(args))
allvars <- args[4:nargs]
nargs2 <- (sum(!is.na(allvars)))
first_vars <- as.character(allvars[1:(nargs2/2)])
second_vars <- as.character(allvars[((nargs2/2)+1):nargs2
#######
combined2 <- merge(fileA, fileB, by.x=c(first_vars),  by.
               all.x=T, all.y=F, sort=F, suffixes = c(".x
…………
```

Might social classifications matter in longitudinal simulations?

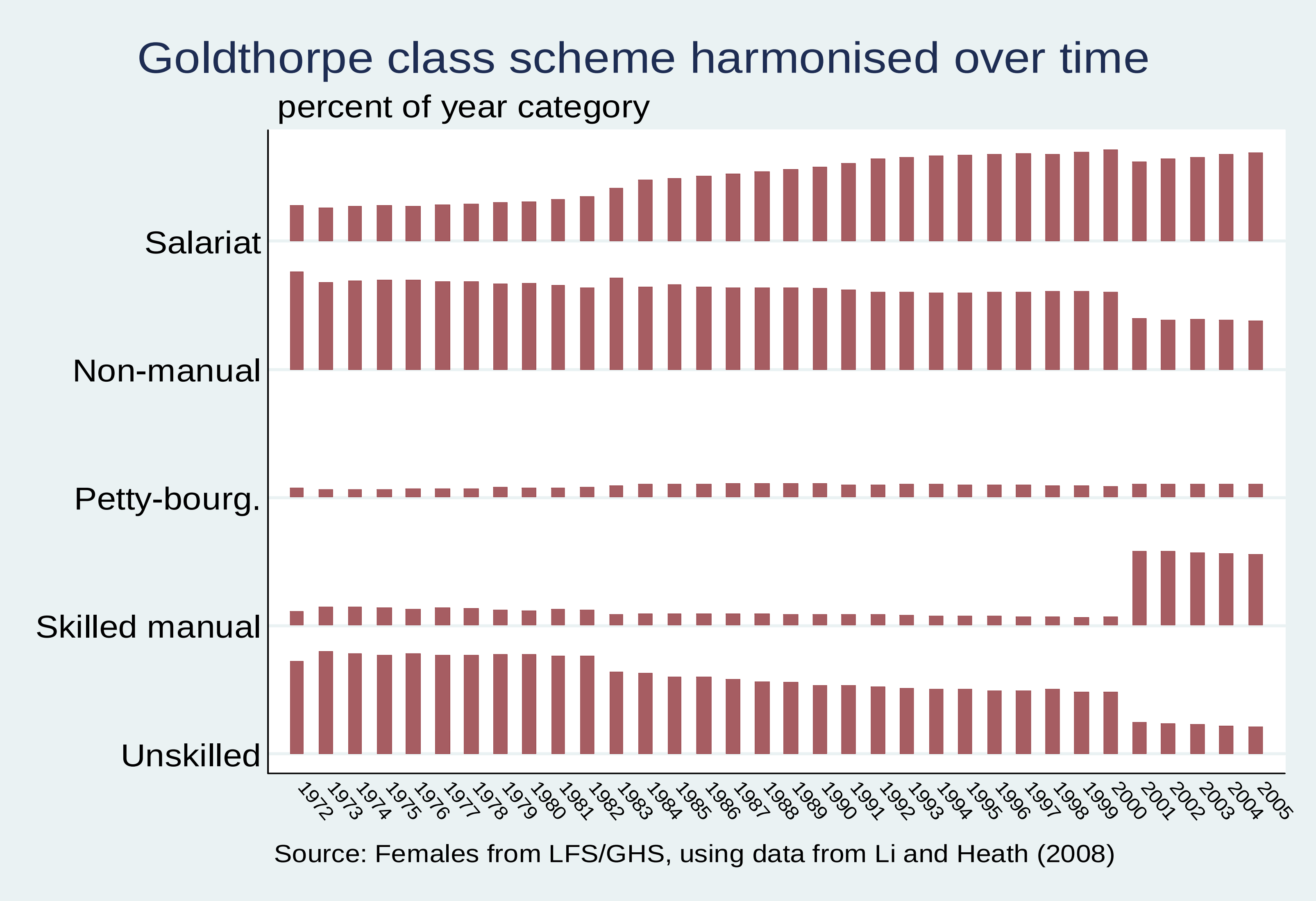
- Things might be pretty robust regardless of measures
- Differences between different measures tend to correlate with age, gender, and change over time
- Differences arising from functional form might be consequential, cf.
  - Crossing a threshold in a {two}-category measure
  - A continuous model without any thresholds

Things happen over time and as people age…



The educ profile represents grade inflation. The income/occ profiles could be one or two things - changing rewards with age; plus or not a general upgrading of rewards across birth cohorts

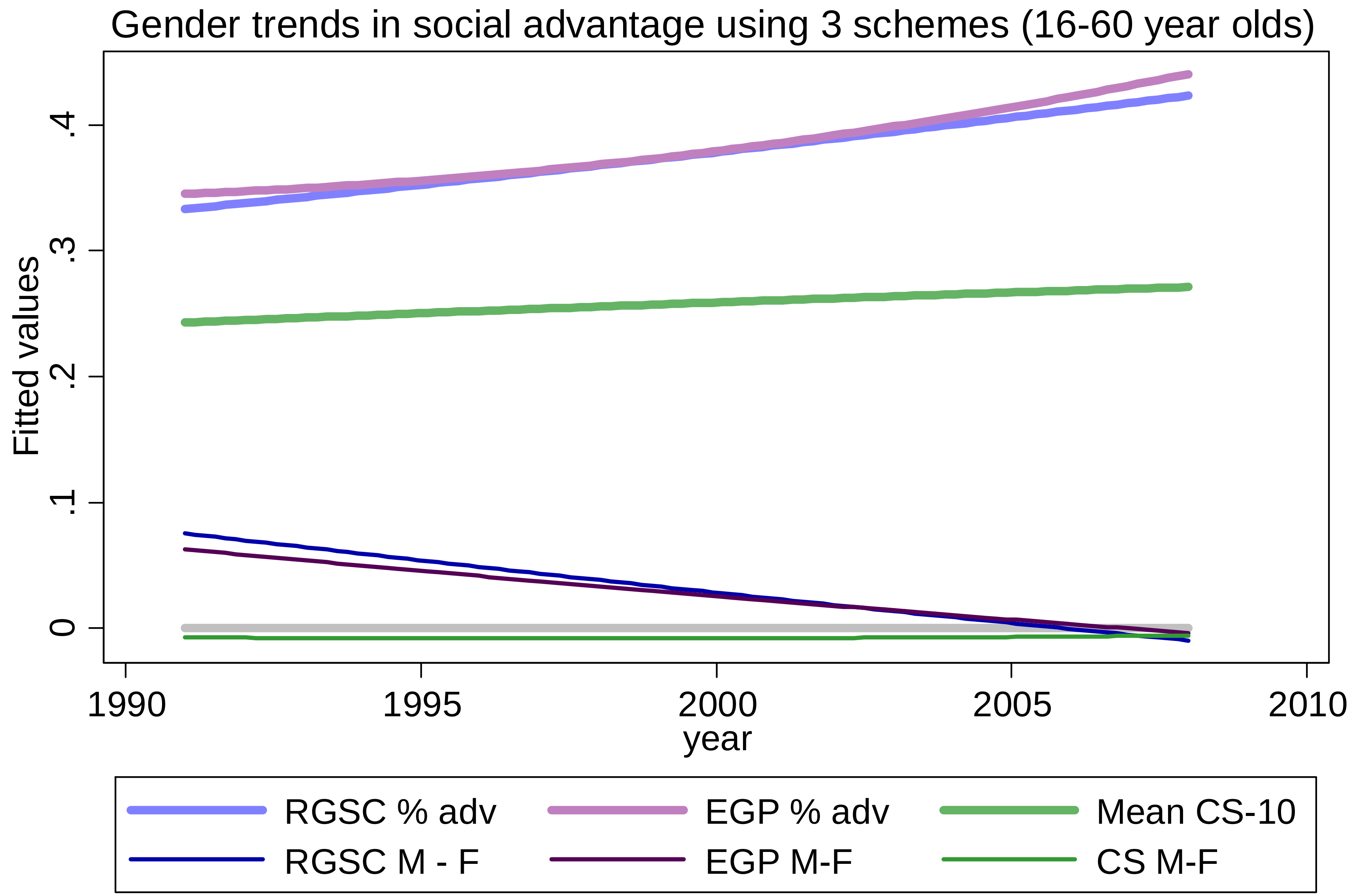
Things happen over time and as people age…



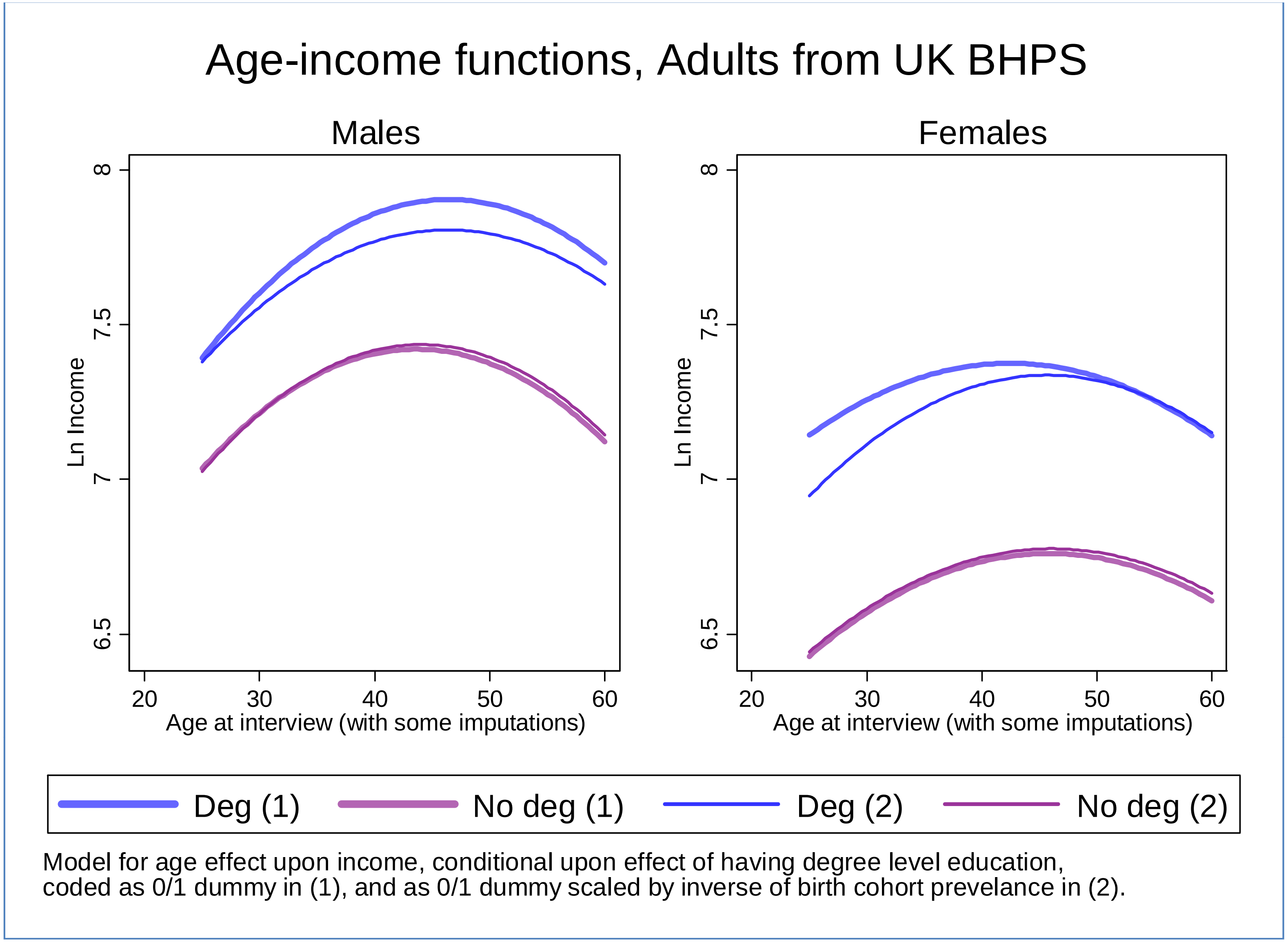
It proves very difficult to separate the experiences of age cohorts from other time trends (gender; industry; administration)

…Different candidate sociological measures routinely place the same people in 'different' positions…

BHPS 1991, binary categorisation of 'advantaged', N=6151						
EGP	**FCAMSIS**	**MCAMSIS**	**NS-SEC**	**RGSC**	**N Male**	**N Female**
I or II	*>66*	*>66*	*1 or 2*	*I or II*		
Y	Y	Y	Y	Y	379	173
Y	N	N	Y	Y	409	289
N	N	N	N	Y	101	90
Y	N	N	Y	N	50	93
Y	Y	N	Y	Y	94	112
Y	N	Y	Y	Y	99	100
					.	.
			All with 1+ records		1371	1047



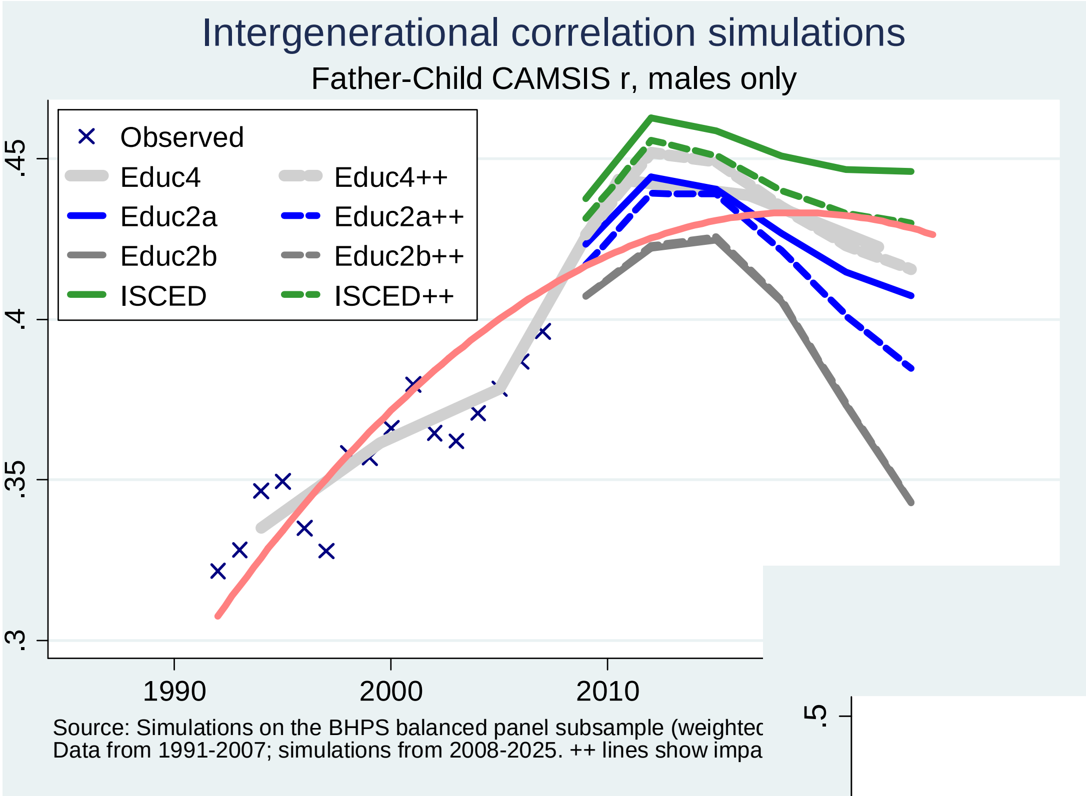
Source: BHPS waves 1-18, x-sectional respondent weighting, all adults aged 16-60 only



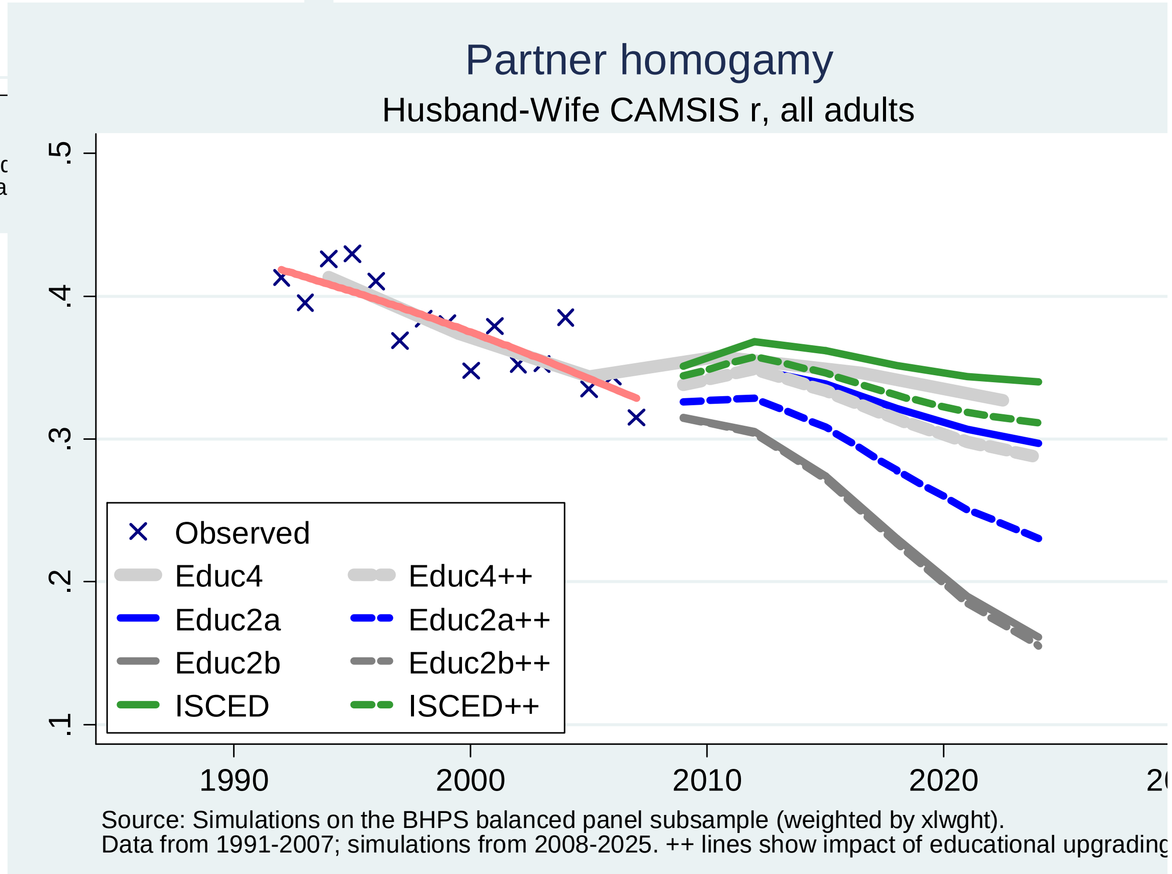
Interpretation: Educ is an important mediator in age-income relationship, but how an education measure is scaled can change interpretation considerably.

Data management in NeISS

- Idea of an infrastructure supporting re-specification of analysis with adjustments to data and models options
    - A set of 'blessed scripts' (which run simulations in target areas subject to X variables), featuring options on them allowing different input data, simulation parameters, and *variable operationalisations*
    - 'Data fusion tool' to facilitate data storage and linking
    - Workflow tools to coordinate tasks (see also MP's talk)
    - *For discussion - something else needed?*



Initial results: Different measures sure as hell do matter to the simulation!



(3) Specification and results from the ageing and inequality microsimulation

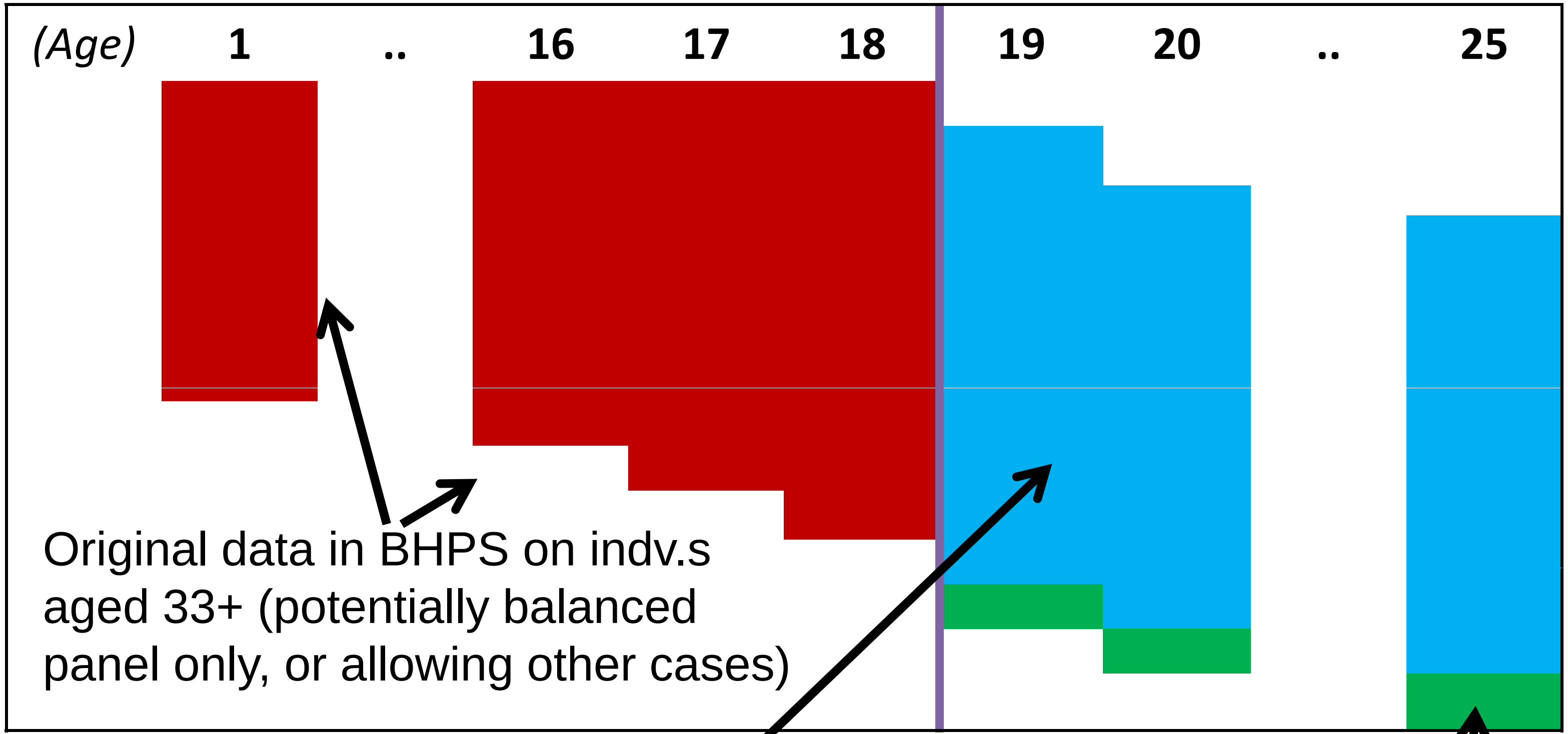
- Use BHPS as baseline resource, in order to calculate year-on-year* probabilities or rules** from one situation to another
- Probabilities are then applied successively to a baseline dataset, projected forward over time, and that data is then summarised (the simulation)

  (with rules applied to ensure that predicted future values are within a a plausible range)

** Have tried either annual updates, or 3-yearly updates, but no great difference so far*

*** Probability = predicted value from a regression model using current predictors plus being informed by data on an individual case's history; 'rule' = a priori condition applied to future value.*

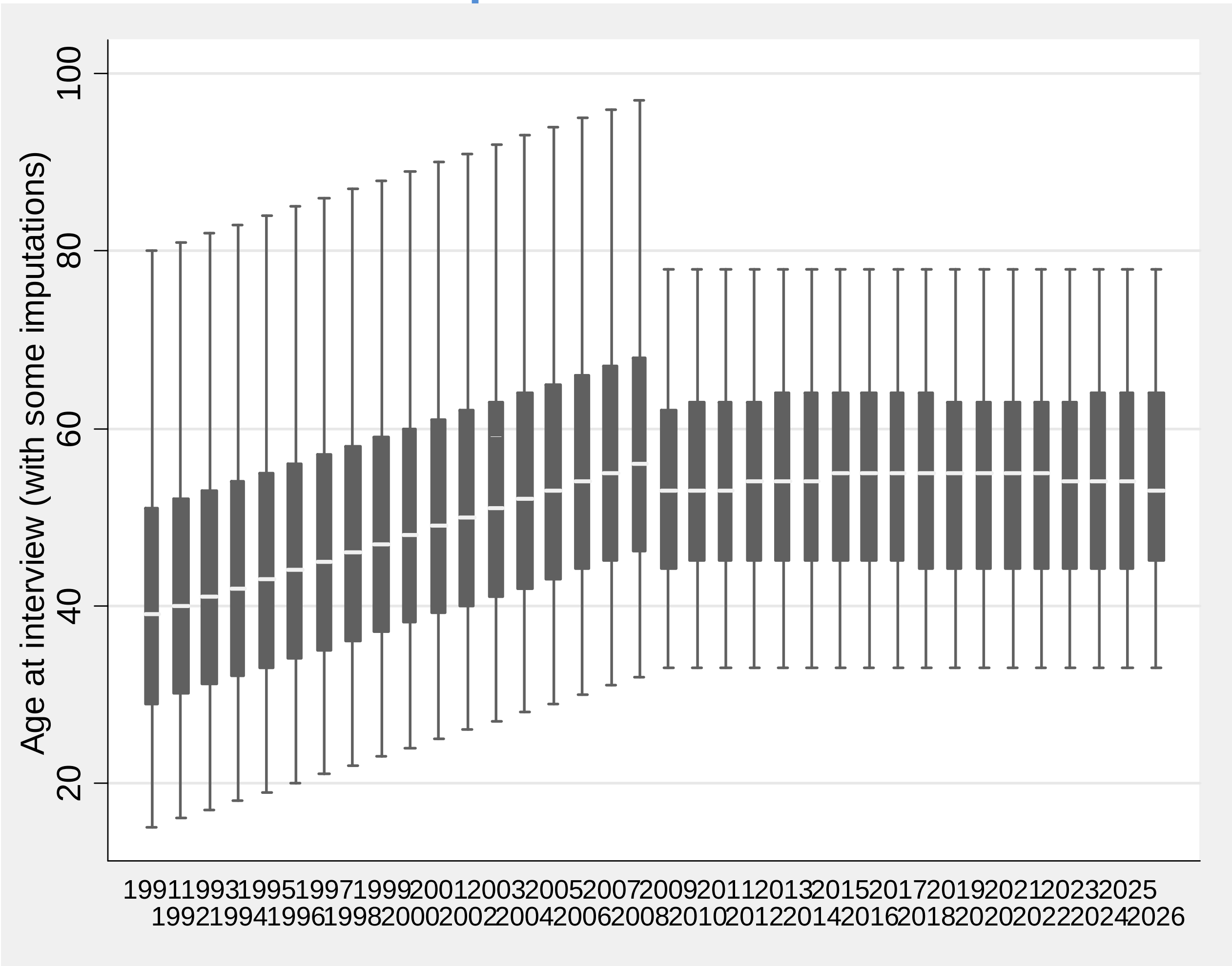
..BHPS simulated data..



Original BHPS people carried forward in time) (if agei <= deathi)

Annual refresher sample of 'young' adults (=replication of adults c33yrs from earlier waves)

Balanced panel source:



Unbalanced panel:

```
. tab year if ~missing(zjbcssm)
```

year	Freq.	Percent
1991	6,151	1.83
1992	5,816	1.73
1993	5,540	1.65
1994	5,772	1.72
1995	5,723	1.70
1996	5,954	1.77
1997	6,932	2.06
1998	6,960	2.07
1999	9,534	2.84
2000	9,730	2.90
2001	9,887	2.94
2002	9,217	2.74
2003	9,285	2.76
2004	9,503	2.83
2005	9,386	2.79
2006	9,780	2.91
2007	9,833	2.93
2008	10,346	3.08
2009	10,591	3.15
2010	10,732	3.20
2011	10,911	3.25
2012	11,039	3.29
2013	11,112	3.31
2014	11,245	3.35
2015	11,320	3.37
2016	11,308	3.37
2017	11,339	3.38
2018	11,179	3.33
2019	10,978	3.27
2020	10,533	3.14
2021	10,332	3.08
2022	10,127	3.02
2023	9,922	2.95
2024	9,726	2.90
2025	9,424	2.81
2026	8,674	2.58
Total	335,841	100.00

Kill some people: drop if agei <= deathi
Make others anew: append using youth_duplicates.dta

Simulating ageing and inequality in Stata

1. Compile microdata long-format panel data extract (typically w1-18 or w1-8)

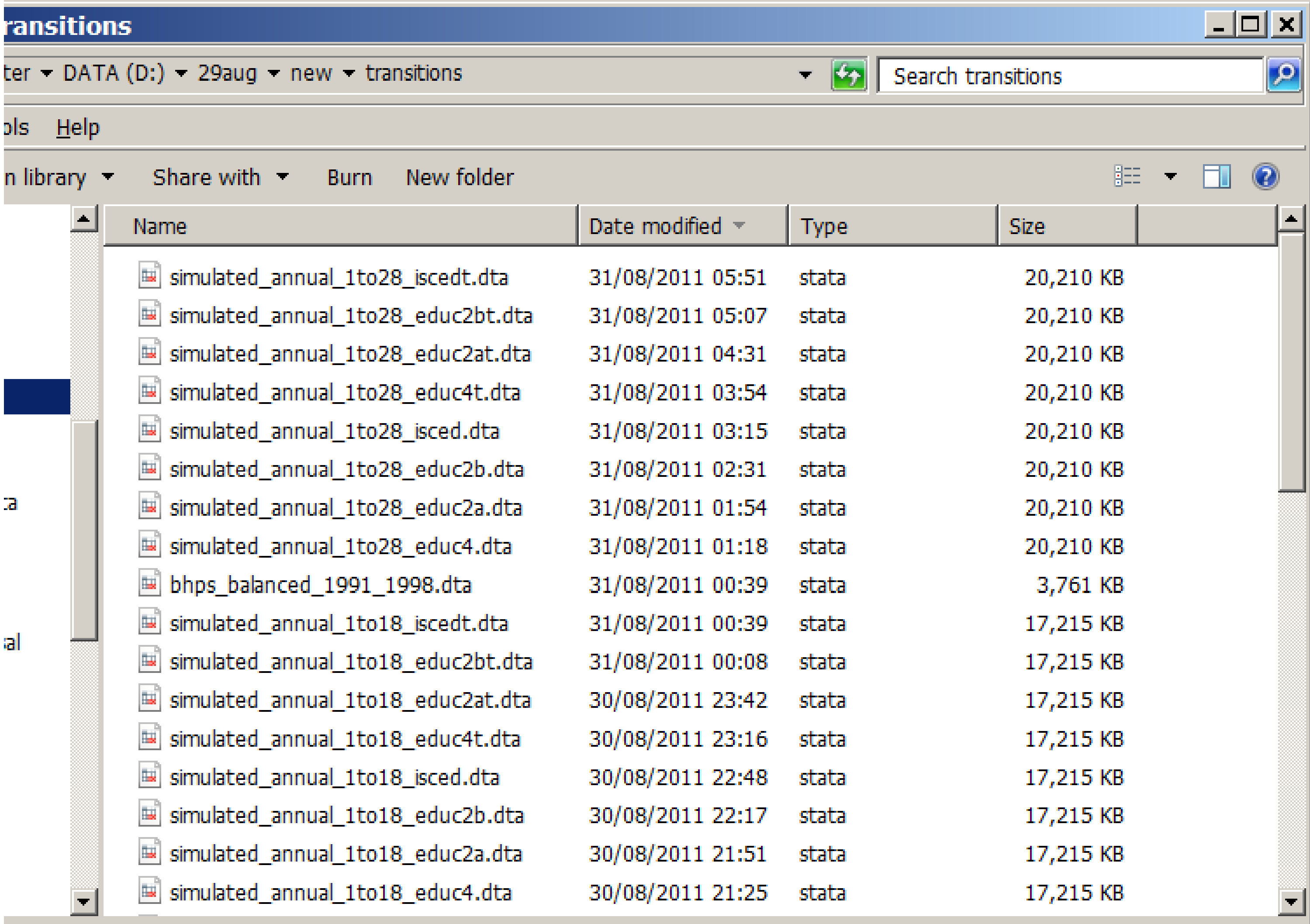
2. Derive alternative measurements of crucial explanatory factors (e.g. education in 4 alternative formats)

3. Simulate forward in time, time-period by time-period, for explanatory variables and linear outcomes

4. Derive summary statistics (statsby - time) from real and simulated data

5. Spend ages trying to get a nice graph of the results

NeISS: (1) & (2) can be re-specified/adapted (arguments in do file); (3), (4) and (5) are pre-specified ('blessed scripts')



An infrastructural resource could help:

- Large datasets; many permutations; slow computations; complex outputs

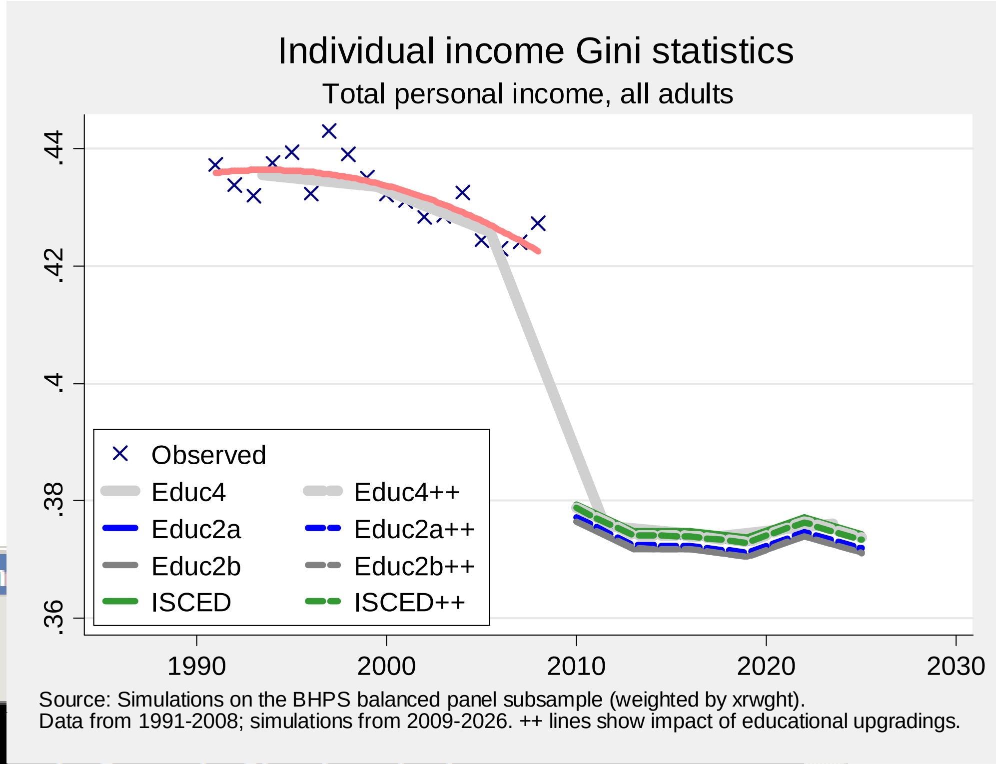
EditPad Pro

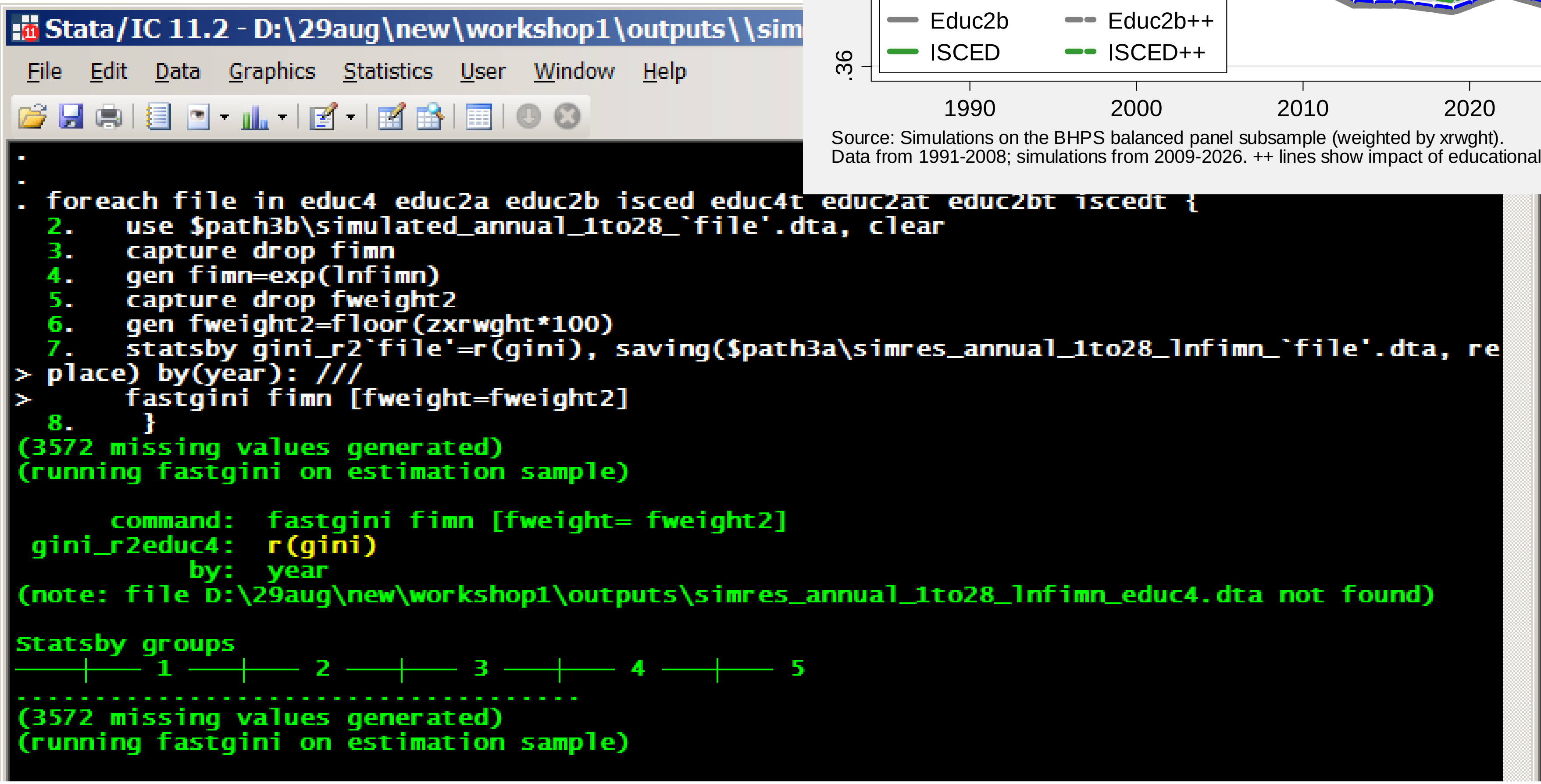
File Edit Project Block Bookmark Tools Extra Convert Options View Help

sim_macros_2.do | simulations_lnfimn_1.do | sims_1_18b.do | sims_1_28b.do

```
program define sim_lin2
 tab $time
 append using ${path9}\${val}.dta /* Add's an extra year's worth of cases */
 append using ${path9}\${val}new.dta /* Add's an input designed to represent young influx (witl
 summarize
 gsort +pid +$time
 tab famtyp1, missing /* Modal family type assignment (these could be modelled, but not done he
 replace famtyp1=famtyp1[_n-1] if (famtyp1==.) & pid==pid[_n-1]
 replace famtyp1=2 if famtyp1[_n-1]==1 & pid==pid[_n-1] & agei>=25 & $time > $endy
 replace famtyp1=3 if famtyp1[_n-1]==2 & pid==pid[_n-1] & agei>=30 & $time > $endy
 replace famtyp1=2 if famtyp1[_n-1]==3 & pid==pid[_n-1] & agei>=50 & $time > $endy
 replace $educ= $educ[_n-1] if pid==pid[_n-1] & missing($educ) /* Takes last year's education :
 codebook $varused, compact /* Outcome variables from original data */
foreach linout in $varsused {
 capture drop sim_val
 gen sim_val=.n
 replace sim_val=`linout' if ~missing(`linout') /* for existing cases, the previous value is us
 capture drop lagout
 gen lagout=`linout'[_n-1] if pid==pid[_n-1]
 capture drop meanout
 egen meanout=mean(`linout'), by(pid)
 replace meanout=meanout[_n-1] if pid==pid[_n-1] & missing(meanout)
 xi: regress `linout' i.sex agei age2 dobyi i.$educ i.$educ*dobyi lagout meanout i.famtyp
 capture drop pred
 predict pred , xb /* Next year's data has a value here because it has non-missing lag and mear
 replace sim_val=pred if sim_val==.n & pred >= -100 & pred <= 100
 capture drop maxyear
 egen maxyear=max($time), by(pid)
 replace `linout'=sim_val if $time==maxyear & maxyear >= $endy /* for new cases, the regressior
 codebook `linout' sim_val , compact /* Outcome variables after adding in last years data */
   }
end
```

Want to redo analysis many times…



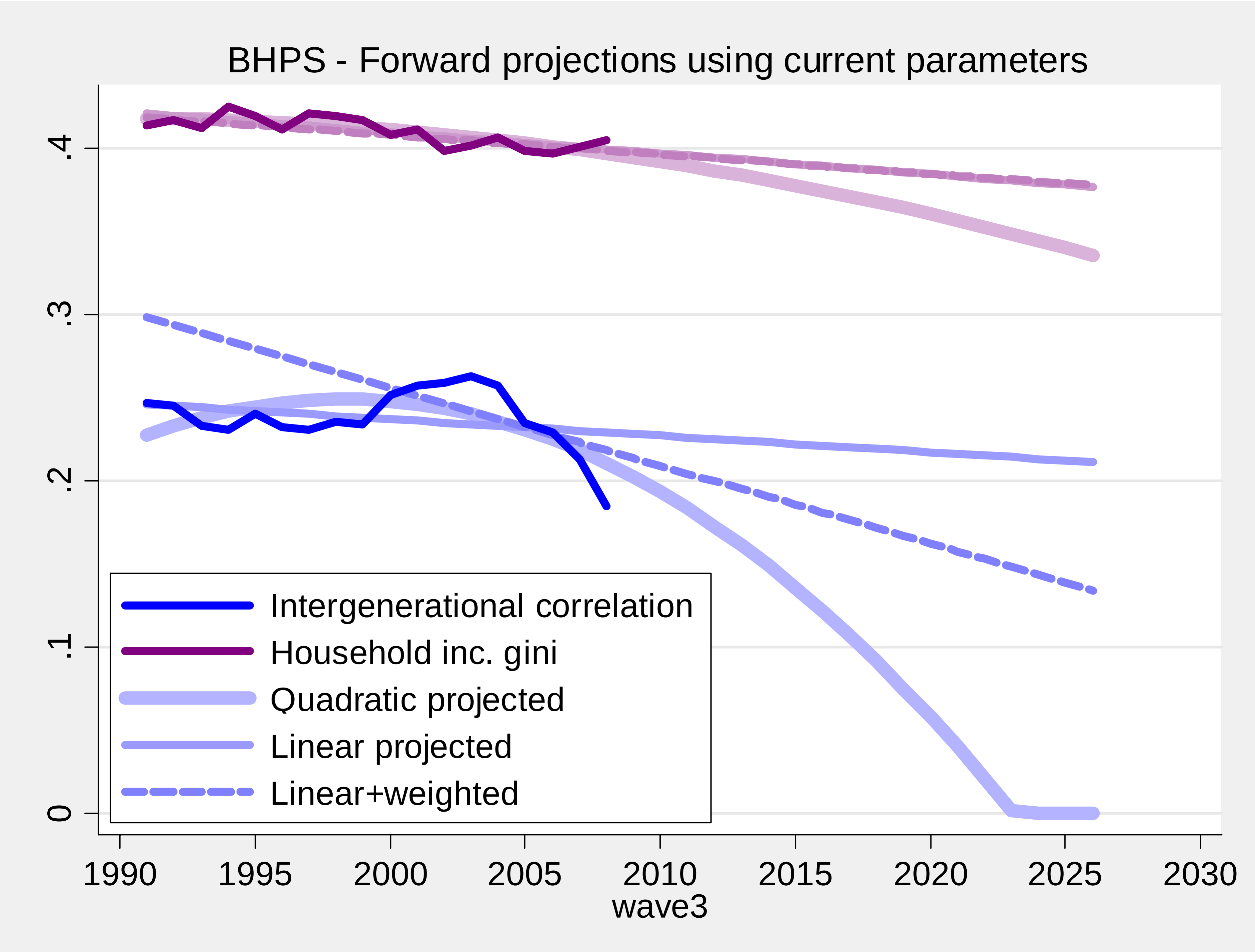


Initial micro-simulation models on BHPS…

- Regression model of outcomes predicted by soci-demographic and socio-economic factors, and lag outcomes, with additional rules or stipulations
  - Rules for continuity/change in family type (age*educ=modal status); and education (stability)
  - Stipulate increase in educational provision; decrease in industrial composition
  - Could consider numerous additional specifications (e.g. transitions responsive to industrial composition)
- Weighted distributional summaries carrying forward last year's cross-sectional weights

Outcomes - a selection of possible measures of social inequality using BHPS 1991-2008

	1991	Mean	2008	R2 with year^2
Intergenerational correlation (CAMSIS) (all adults)	25	24	18	37
Intergenerational correlation (CAMSIS) (men only)	25	24	19	32
Husband-wife homogamy (CAMSIS r)	39	38	32	67
Personal income Gini (all adults)	45	44	43	87
Personal income Gini (men only)	40	40	41	46
Household income Gini (all adults)	41	41	40	63
Occupational Gini (CAMSIS) (all adults)	29	28	27	93
Household occupational Gini (CAMSIS)	32	31	29	86
Percent of all adults in EGP I	11	13	16	84
Percent of all adults in EGP I or II	28	32	40	96
Percent of all adults in RGSC I or II	28	32	38	96
Source: BHPS cross-sectional aggregates, weighted using {w}xrwght.				

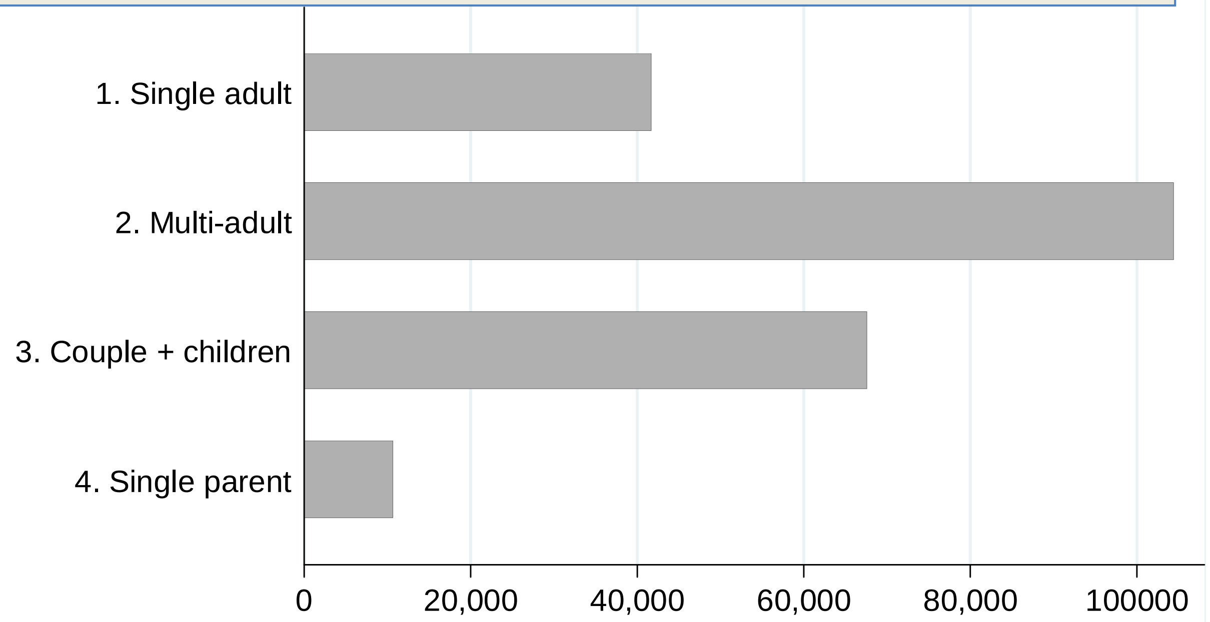

BHPS - Forward projections using current parameters
0
.1
.2
.3
.4
1990
1995
2000
2005
2010
2015
2020
2025
2030
wave3
Intergenerational correlation
Household inc. gini
Quadratic projected
Linear projected
Linear+weighted

Variations in deterministic parameters

- Here we'll include in models the influences of educational level and family type (and artificially adjust educational qualifications' prevalence by age cohort)

Variable	Obs	Unique	Mean	Min	Max	Label
isced	210571	8	20.72271	12	32	
zqfedhi	209785	12	6.957461	1	12	highest educational qualification
educ4	209785	4	2.808118	1	4	BHPS 4-fold educational level classification
educ2a	209785	2	.1169197	0	1	Degree education or above
educ2b	209785	2	.2787092	0	1	Low school level or below

..Many more variations of these and other measures are possible, for future consideration…



Methodological topics

– Comparison between analyses which use different measures of position within the inequality structure

e.g. occupations; education; income; wealth

– Model of the feedback effects on those positions of trends in national and local distributions

variously measured

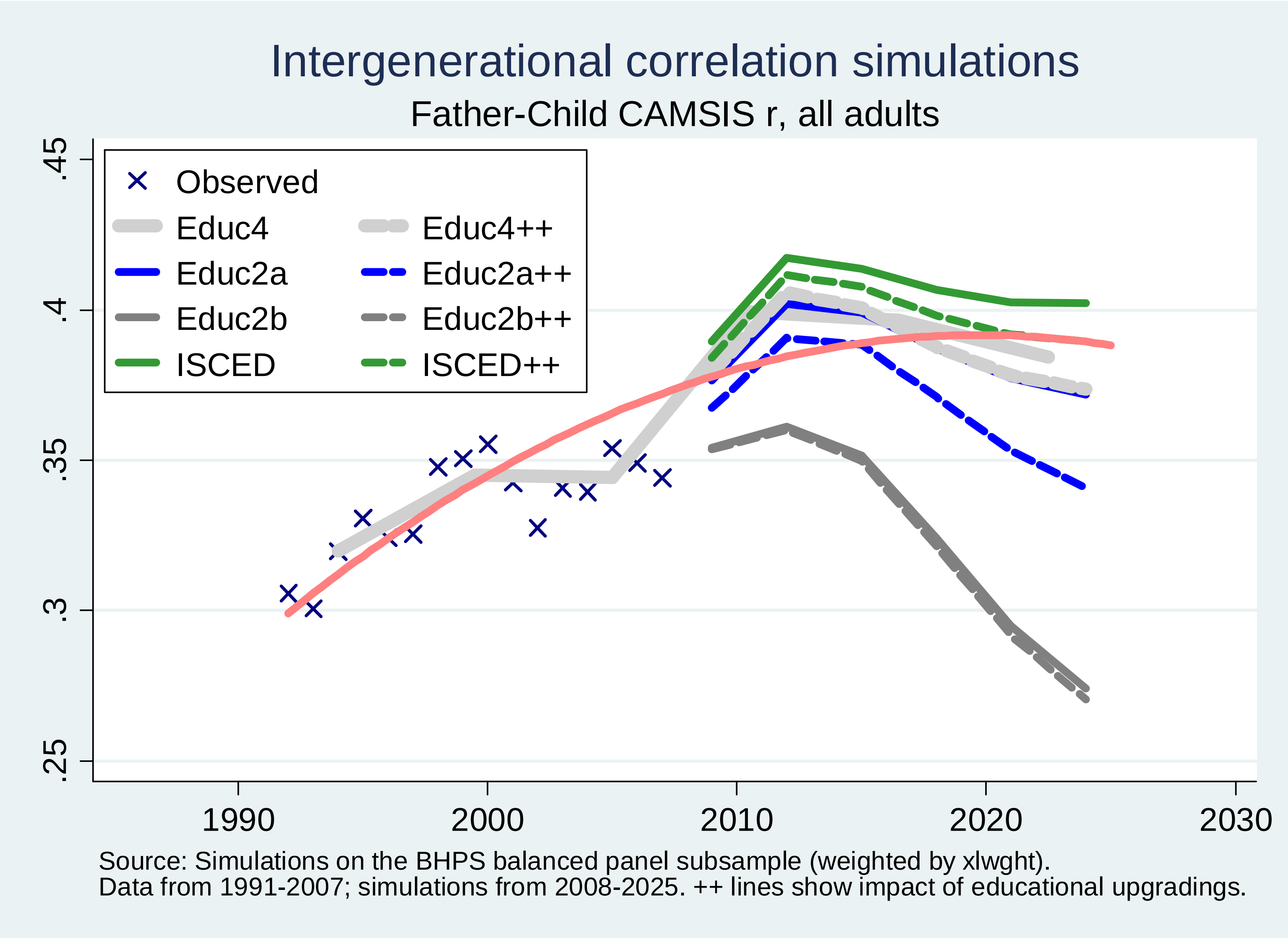
– Modelling of the feedback effects of trends in national and local demographics (e.g. family structures; immigration)

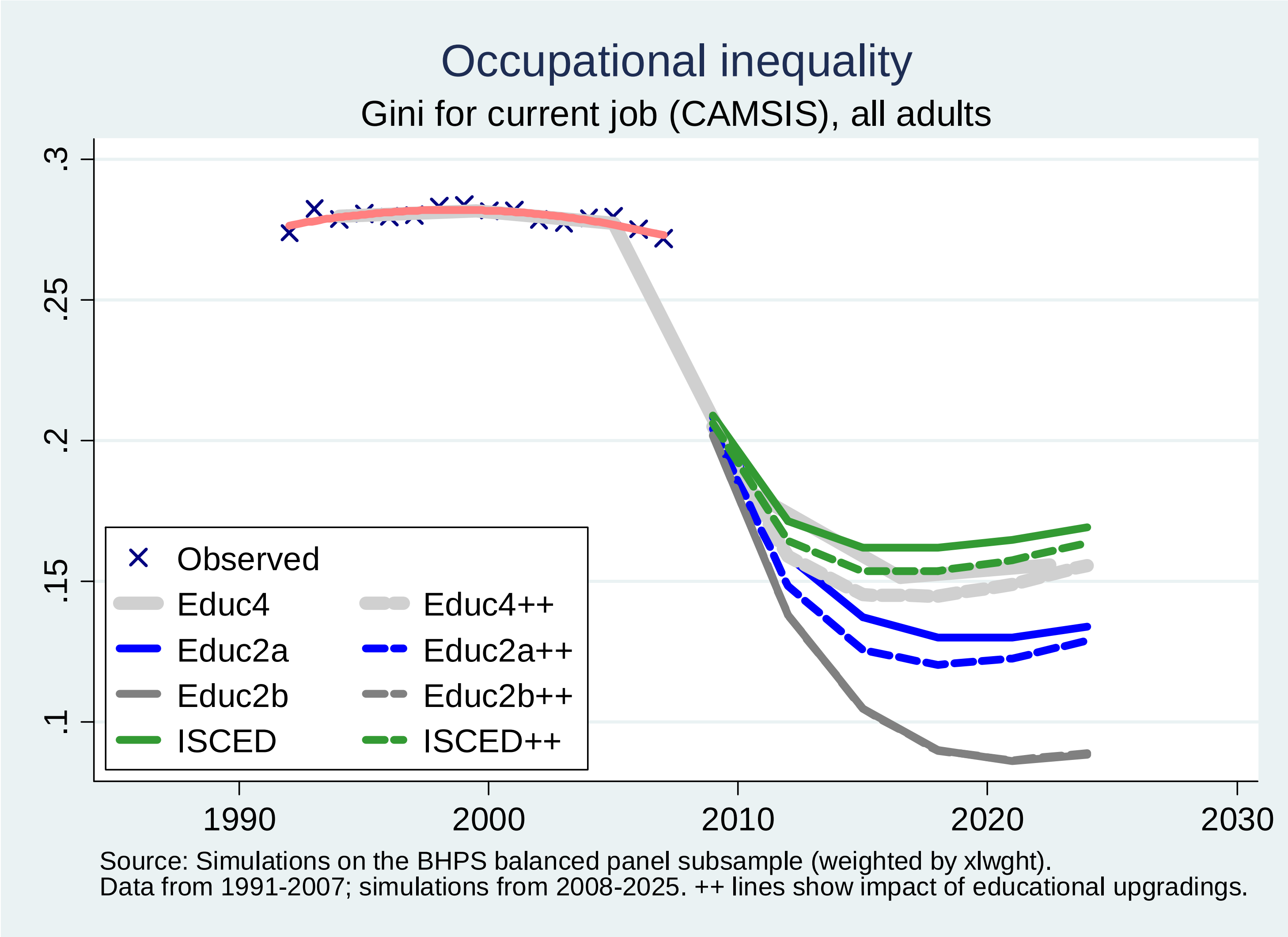
variously measured

In the following application…

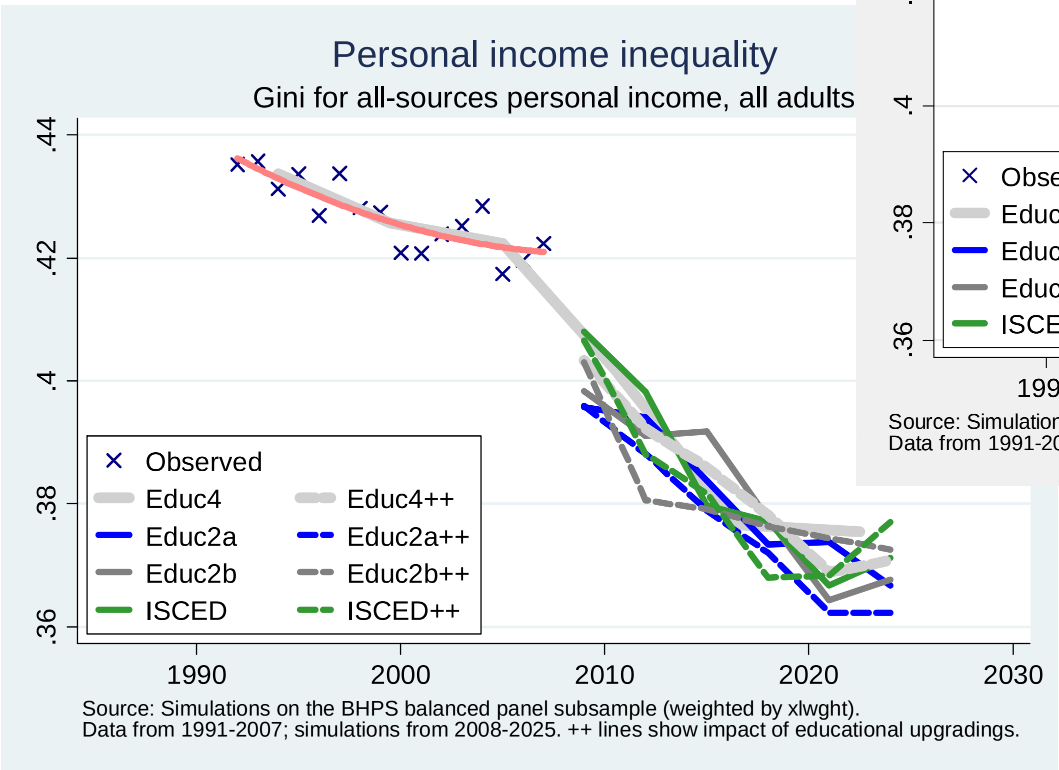
- BHPS balanced and unbalanced panel
  - (carry forward 1991-2008 respondents every year till 2025)
- Predict next year's outcome from predicted values of a regression with explanatory variables of the current outcome (observed or simulated), plus gender, age, dob, educational level, family type, and age*educ interaction
- Shortcuts: global imputation for family type; ignoring spouse's changes; …
- Variable parameters summarized below:
  - 4 different education measures
  - 4 different treatments (increasing educ for later cohorts only)

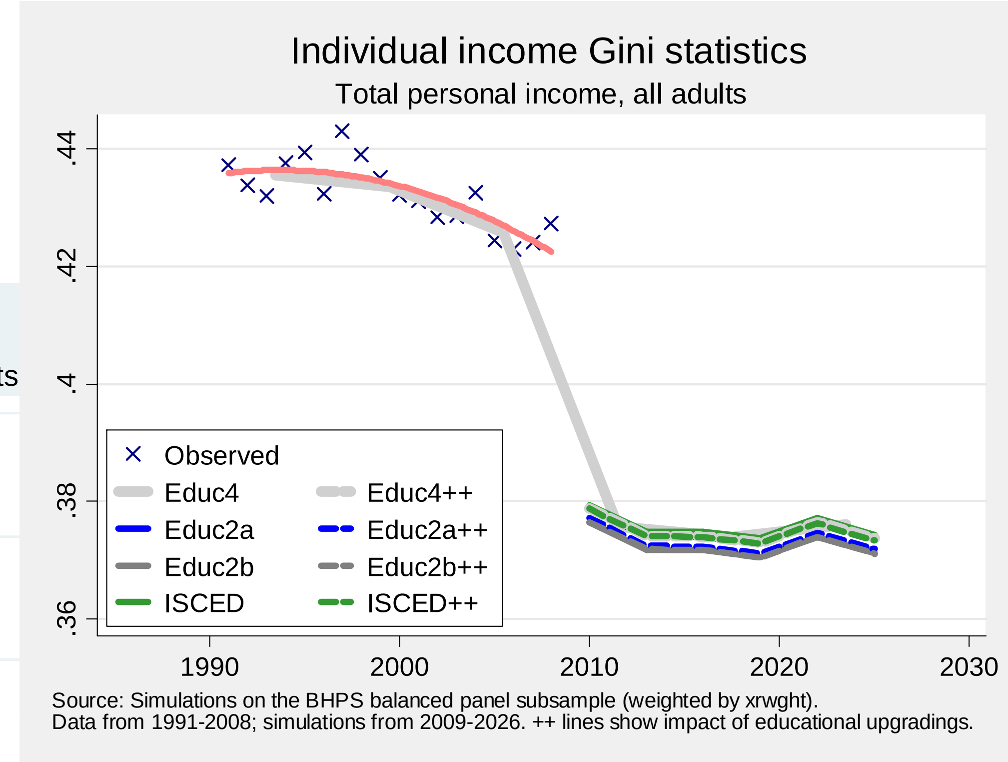
First evidence on the effects of different sociological classifications



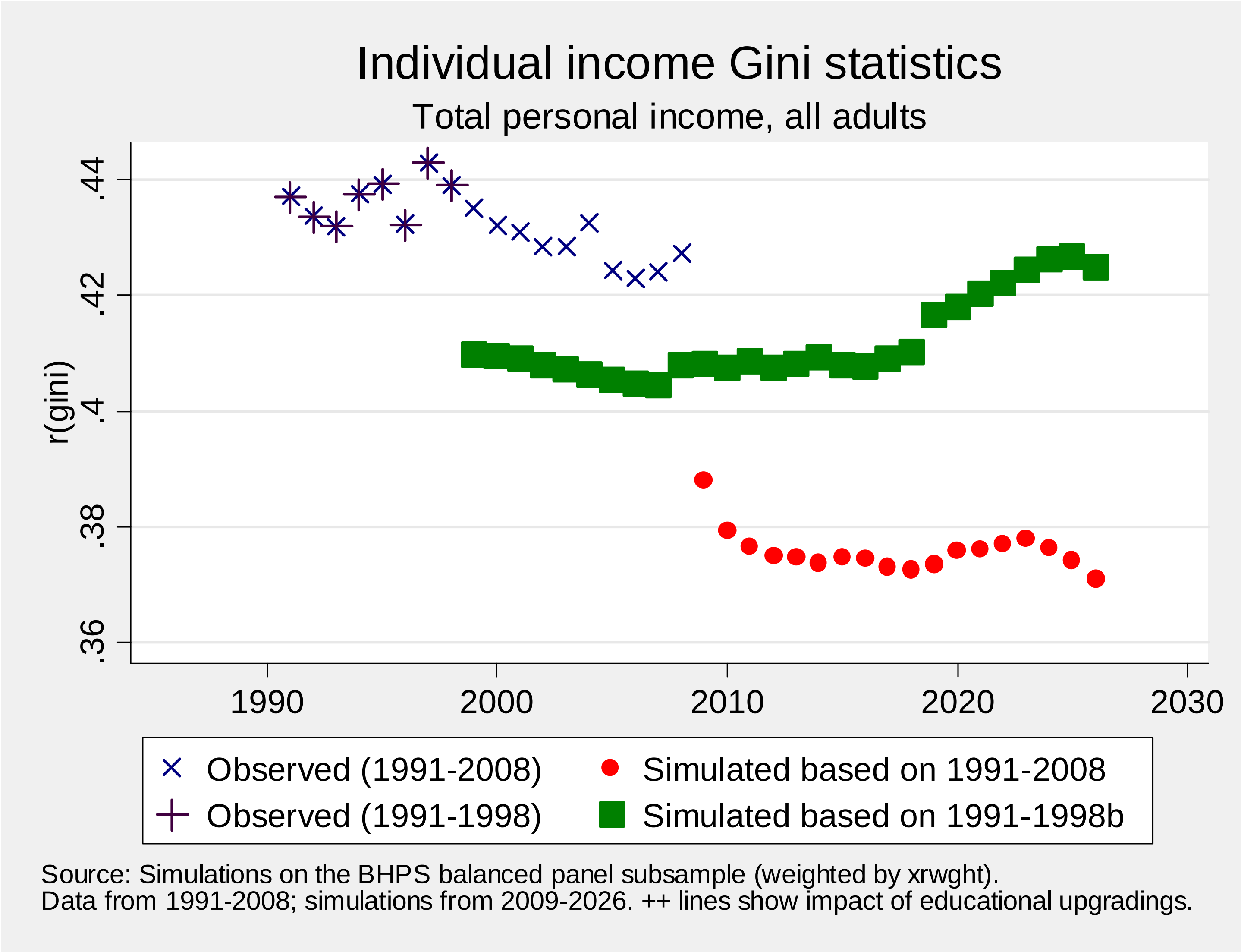


Gini calculations on income and occupations: so far the regression model generating the simulated values isn't doing a good job of summarising inequality (as it tends to reduce heterogeniety between individuals)





- When greater within-person dependence is used, however, the variable operationalisation impacts diminish



Testing on real data reveals numerous problems…

- Regression model irons out heterogeneity
- Profile influenced by observed profile

Conclusions

- Simulations and social classifications
  - Wealth of social classifications metadata (www.dames.org.uk)
  - Simulations offer a tool for evaluating classifications (haven't previously been used for this?)
  - Classification permutations offer new alternatives to the simulation communities
  - {NeISS role in infrastructural support}

Preliminary findings suggest:

  - Measures are important – differences between measures can matter a lot, and they can matter more than do differences between treatments!
  - Gaps open up: Longer-term longitudinal trends susceptible to differences in measures

More work needed

- Improve the simulation model
  - Allow non-linear outcomes
  - Models potentially enhanced by other supplementary micro-data, e.g. on transitions between rarer states (cf. Zaida et al. 2009b)
  - Better accounts of individuals in their households
  - Evaluation of simulation using existing data
- Building the simulation infrastructure tool
  - » Taverna workflows
  - » File storage
  - » We need users, contacts, and application exemplars

References

- Birkin, M., Procter, R., Allan, R., Bechhofer, S., Buchan, I., Goble, C., et al. (2010). The elements of a computational infrastructure for social simulation. *Philosophical Transactions of the Royal Society, Series A, 368(1925), 3797-3812.*
- Gilbert, G. N. (2008). *Agent-Based Models. Thousand Oaks, Ca.: Sage.*
- Gilbert, G. N., & Troitzsch, K. G. (2005). *Simulation for the Social Scientist, 2nd Edition. Maidenhead, Berkshire: Open University Press.*
- O'Donoghue, C., Leach, R.H., & Hynes, S. (2009) "Simulating earnings in dynamic microsimulation models", in Zaidi, A., Harding, A., & Williamson, P. (Eds.). (2009a). *New Frontiers in Microsimulation Modelling. Farnham: Ashgate, pp381-412.*
- Zaidi, A., Evandrou, M., Falkingham, J., Johnson, P., & Scott, A. (2009b) "Employment transitions and earnings dynnamics in the SAGE model", in Zaidi, A., Harding, A., & Williamson, P. (Eds.). (2009a). *New Frontiers in Microsimulation Modelling. Farnham: Ashgate, pp351-379.*
- Zaidi, A., Harding, A., & Williamson, P. (Eds.). (2009a). *New Frontiers in Microsimulation Modelling. Farnham: Ashgate.*

- Simulation models can be used to **project over time** in order to estimate emergent social-structural patterns. The NeISS project (National e-Infrastructure for Social Simulation, www.neiss.org.uk) is a UK initiative in supporting the construction, estimation and interpretation of social simulation models applied to a variety of scenarios. In this paper, I will present results from one of the exemplar projects within NeISS, an analysis of **'ageing and inequality'**, which is designed to **model the development of social inequality over time in response to trends in major socio-demographic and socio-economic changes** (such as the aging population, changing family formation patterns, changing patterns in educational provision, and changing occupational/industrial opportunity structures). Social inequality indicators used include measures of income inequality, occupational inequality, and social mobility. The data is initially parameterised around annual transition patterns in contemporary Britain, though it should in principle be generalisable to other data scenarios. A unique contribution of the NeISS project is its capacity to support multiple replications of simulations using different underlying measurement instruments of the same concepts – in this paper, we explore the impact of different approaches to measuring occupational circumstances, educational attainment and ethnicity in the context of the simulation model.

…from the NeISS application…

- "The key substantive question concerns the interaction between ageing effects and [the] nature and impact of socio-economic inequalities. These issues involve complex, non-linear processes that are suited to simulation approaches. The exemplar will enable study of the impact of alternative socio-economic measures and resources within a micro-level simulation analysis of socio-economic inequalities across age groups, premised upon large scale social survey data (such as British Household Panel Survey, Labour Force Survey, General Household Survey and UK Census based data)" (WP 4.1.4)