Stefan Bach

Distributional analyses and microsimulation modelling based on integrated data from household surveys and income tax files

EUROMOD WORKSHOP – 2019
Milano, September 25, 2019

Outline

Data integration: Survey data and income tax files

The entire income distribution in Germany

Tax distribution: Who bears the tax burden in Germany?

Prospect: DINA Germany

Why data integration?

Survey data on household income and wealth

Cover the entire population, designed for socio economic research

- Germany: SOEP, Income and consumption survey, EU Silc, HFCS

Drawbacks for distributional analyses

- Underrepresentation of relevant groups, e.g. rich households
  - Sampling error due to small sample size
  - Selectivity (“non response bias“, “middle class bias“)

Drawbacks for microsimulation

- Missing variables relevant for taxation
  - e.g. specific incomes such as capital gains, income related expenses, other deductions

Administrative data: personal income tax files

Full representation of characteristics relevant for taxation

- Includes high-income households
- No measurement error

Drawbacks for distributional analyses and microsimulation

- Does not include non-filers, especially low-income households
- Does not include corporate income
- Only includes information relevant and disclosed for taxation
  - Tax avoidance and evasion
- Limited information on socio economic background
- Legal restrictions on micro data access: Data protection and tax secrecy laws

Data integration: impute or match?

Integration of micro data

Congruent variables in both datasets

- income groups, gender, family status, number of children, age groups, regions, employment status, etc.
- Conditional independence assumption with respect to congruent variables

Regression imputation

- Econometric estimation based on source data set
- Prediction into target data set
- E.g. for single characteristics

Statistical matching procedures

- Matching of similar observations according to score based on congruent variables
- Maintaining correlation structures of imputed variables

Integration of aggregated data

Editing both datasets by pre-defined cells

- e.g. by income groups, age groups and gender

The entire income distribution in Germany

Integrating income tax files and SOEP 1992-2005

Income distribution up to the very top

Effective income tax rates, optimal top tax rate

Publications

Bach, Corneo, Steiner (2009): From Bottom to Top: The Entire Income Distribution in Germany, 1992 - 2003. Review of Income and Wealth 55 (2), 331-359.

Bach, Corneo, Steiner (2013): Effective Taxation of Top Incomes in Germany. German Economic Review 14 (2), 115-137.

Bach, Corneo, Steiner (2012): Optimal top marginal tax rates under income splitting for couples. European Economic Review 56 (6), 1055-1069.

Empirical Approach

Thoroughly editing economic income in both data sets

Income relevant for income tax assessment

Editing SOEP data at the taxpayer level

Potential tax units

Statistical match of SOEP potential taxpayers to income tax files

Constrained matching within pre-defined cells

- gross income groups and marital status

Data integration strategy

Constrained statistical matching

Within predefined cells: gross income groups and marital status

Conditionally on a number of common variables

- gross income, main income source, occupational status, marital status, age group, family type, number of children,
  other tax-relevant information

Maintaining the weighted distribution of both data sets

Using LP optimization routines

Network simplex algorithm (transportation problem)

- performed by ilog CPLEX and implemented in JAVA
- SOEP data: donor (supplier), income tax file data: host (demander)

Distance measure: absolute deviation between normalized common variables

Resulting integrated data base

Bottom parts of income distribution: SOEP aggregates > income tax file aggregates

- Nonfilers left behind

Top parts of income distribution: income tax file aggregates > SOEP aggregates

Data matching methodology

Statistical matching

Nearest neighbor matching

Conditional **I**ndependence **A**ssumption

Constrained matching approach

Each observation (record) of SOEP is matched to a certain number of records in the income tax files

Correlation structure between the common matching variables and other variables is maintained

LP transportation model

Records of data set A (B) as supply (demand) nodes

Survey weights w_{ij} of A and B as volumes supplied (demanded) by each A (B) record

Distance measure d_{ij}, as the costs of shipped goods between A and B

Minimize the weighted costs over all data records (n_A, n_B)

- restriction: for each record the weighted number of cases matched from A to B equals the sum of weights in the respective data set

$$\min \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij} w_{ij}$$

$$s.t. \quad \sum_{j=1}^{n_B} w_{ij} = w_i, \forall i, \quad \sum_{i=1}^{n_A} w_{ij} = w_j, \forall j, \quad w_{ij} \geq 0, \ \forall i, j$$

Distance measure: e.g. absolute deviation, Euclidian, or Mahalanobis distance

- absolute deviation after normalizing all variables

Distribution of gross market income in Germany, 1992-2007

	Gross market income[1], capital gains excluded							1992 = 100					
	1992	1995	1998	2001	2003	2005	2007	1995	1998	2001	2003	2005	2007
Average income at 2000 prices[2]													
Mean income (Euro)	19 960	19 705	19 823	19 584	19 135	18 502	19 098	98.7	99.3	98.1	95.9	92.7	95.7
Median income (Euro)	12 494	11 332	9 722	8 219	7 024	5 670	6 729	90.7	77.8	65.8	56.2	45.4	53.9
Relative difference[3] (%)	46.8	55.3	71.2	86.8	100.2	118.3	104.3	118.1	152.1	185.3	213.9	252.5	222.7
Gini coefficient[4]	0.6155	0.6209	0.6389	0.6538	0.6617	0.6835	0.6832	100.9	103.8	106.2	107.5	111.1	111.0
Generalized entropy measures[4] [5]													
GE(0)	1.9406	2.0131	2.1834	2.2217	2.2737	2.3855	2.2695	103.7	112.5	114.5	117.2	122.9	116.9
GE(1)	0.7810	0.7868	0.8472	0.8905	0.9025	1.0033	1.0353	100.7	108.5	114.0	115.5	128.5	132.6
GE(2)	4.3527	5.4620	7.3885	8.7560	18.7362	30.5886	55.8182	125.5	169.7	201.2	430.4	702.7	1 282.4
Structure in % by income fractiles													
1st decile	- 0.83	- 0.96	- 0.95	- 0.88	- 0.67	- 0.53	- 0.44	115.4	113.8	105.8	80.0	63.9	52.8
2nd decile	0.05	0.04	0.03	0.03	0.03	0.03	0.05	86.9	69.7	68.3	64.9	59.1	100.1
3rd decile	0.19	0.16	0.13	0.13	0.11	0.09	0.13	86.4	66.7	70.8	57.6	49.0	67.7
4th decile	1.18	0.98	0.72	0.66	0.54	0.39	0.36	83.1	60.8	55.4	45.3	32.7	30.0
5th decile	4.24	3.67	3.02	2.57	2.26	1.81	1.97	86.5	71.1	60.5	53.3	42.8	46.5
6th decile	8.23	8.12	7.46	6.61	6.01	5.23	5.61	98.6	90.7	80.2	73.0	63.5	68.2
7th decile	12.06	12.34	12.00	11.43	11.13	10.35	10.30	102.3	99.4	94.7	92.2	85.8	85.4
8th decile	15.69	16.08	16.02	16.10	16.33	15.68	15.30	102.5	102.1	102.6	104.1	99.9	97.5
9th decile	20.14	20.51	20.85	21.26	21.99	21.93	21.00	101.8	103.5	105.5	109.2	108.9	104.3
10th decile	39.04	39.06	40.72	42.11	42.28	45.02	45.73	100.1	104.3	107.9	108.3	115.3	117.1
Top 1%	11.23	10.66	11.57	12.22	11.55	13.57	14.76	94.9	103.1	108.8	102.9	120.9	131.5
Top 0.1%	4.19	3.86	4.37	4.67	4.23	5.49	6.25	92.3	104.5	111.6	101.1	131.0	149.3
Top 0.01%	1.63	1.56	1.82	1.96	1.85	2.49	2.94	95.6	112.1	120.3	113.7	153.1	181.1
Top 0.001%	0.55	0.59	0.72	0.77	0.84	1.11	1.34	107.3	130.0	138.9	151.8	199.9	242.4
Top 0.0001%	0.16	0.20	0.24	0.24	0.37	0.45	0.57	125.5	151.8	156.2	237.8	290.2	362.7
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.0	100.0	100.0	100.0	100.0	100.0

1) Income from business activity, wage income, capital income, exclusive public and private pensions; measured at the individual level.- 2) Deflated by consumer price index.- 3) Difference of ln(mean) and ln(median).- 4) In cases with zero or negative income this income is replaced by 1 Euro.- 5) GE(0) is the mean logarithmic deviation, GE(1) is the Theil index, and GE(2) is half the square of the coefficient of variation.
Source: ITR-SOEP data base.

Source: Stefan Bach, Giacomo Corneo, Viktor Steiner (2009): From Bottom to Top: The Entire Income Distribution in Germany, 1992 - 2003. Review of Income and Wealth 55 (2), 331-359.

Distribution of gross market income for individual and pooled income of spouses, 1992-2007

Gross market income[1], capital gains excluded	Separated income of spouses							Pooled income of spouses[2]						
	1992	1995	1998	2001	2003	2005	2007	1992	1995	1998	2001	2003	2005	2007
Average income at 2000 prices[3]														
Mean income (Euro)	19 960	19 705	19 823	19 584	19 144	18 439	19 006	19 960	19 705	19 823	19 584	19 144	18 439	19 006
Median income (Euro)	12 494	11 332	9 722	8 219	7 028	5 651	6 697	17 115	16 796	16 187	14 745	13 443	11 589	12 407
Relative difference[4] (%)	46.8	55.3	71.2	86.8	100.2	118.3	104.3	15.4	16.0	20.3	28.4	35.4	46.4	42.6
	Summary measures of inequality							Summary measures of inequality						
Gini coefficient[5]	0.6155	0.6209	0.6389	0.6538	0.6617	0.6835	0.6832	0.5213	0.5347	0.5570	0.5773	0.5883	0.6141	0.6098
Generalized entropy measures[5) 6)]														
GE(0)	1.9406	2.0131	2.1834	2.2217	2.2737	2.3855	0.5791	1.3656	1.4786	1.6547	1.7139	1.7911	1.8870	1.7201
GE(1)	0.7810	0.7868	0.8472	0.8905	0.9025	1.0033	0.5807	0.5672	0.5862	0.6451	0.6914	0.7087	0.8019	0.8213
GE(2)	4.3527	5.4620	7.3885	8.7560	18.7362	30.5886	29.7012	2.7407	3.5291	4.9548	5.6484	12.3050	19.0999	35.5865
	Structure by income fractiles[7] in percent							Structure by income fractiles[7] in percent						
1st - 5th decile	4.83	3.89	2.95	2.51	2.27	1.79	2.06	20.12	18.28	16.98	15.71	14.87	13.94	14.80
6th - 9th decile	56.13	57.04	56.34	55.38	55.46	53.20	52.21	51.83	53.16	53.06	52.97	53.50	52.06	51.02
10th decile	39.04	39.06	40.72	42.11	42.28	45.02	45.73	28.05	28.56	29.95	31.32	31.63	34.00	34.18
Top 1%	11.23	10.66	11.57	12.22	11.55	13.57	14.76	7.40	7.07	7.76	8.32	7.87	9.37	10.36
Top 0.1%	4.19	3.86	4.37	4.67	4.23	5.49	6.25	2.75	2.57	2.96	3.22	2.92	3.83	4.38
Top 0.01%	1.63	1.56	1.82	1.96	1.85	2.49	2.94	1.07	1.05	1.26	1.36	1.31	1.77	2.11
Top 0.001%	0.55	0.59	0.72	0.77	0.84	1.11	1.34	0.36	0.40	0.50	0.54	0.60	0.78	0.97
Top 0.0001%	0.16	0.20	0.24	0.24	0.37	0.45	0.57	0.10	0.13	0.16	0.16	0.27	0.32	0.42
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

1) Income from business activity, wage income, capital income, exclusive public and private pensions; measured at the individual level.- 2) Married couples: half of the joint income is assigned to each spouse.- 3) Deflated by consumer price index.- 4) Difference of ln(mean) and ln(median).- 5) In cases with zero or negative income this income is replaced by 1 Euro.- 6) GE(0) is the mean logarithmic deviation, GE(1) is the Theil index, and GE(2) is half the square of the coefficient of variation.- 7) Ranking according to gross marked income, separated income of spouses.
Source: ITR-SOEP data base.

Source: Stefan Bach, Giacomo Corneo, Viktor Steiner (2009): From Bottom to Top: The Entire Income Distribution in Germany, 1992 - 2003. Review of Income and Wealth 55 (2), 331-359.

Distribution of gross income and net income, pooled income of spouses, 1992-2005

Gross market income[1], capital gains excluded	Gross income[2], pooled income of spouses[3]						Net income[4], pooled income of spouses[3]					
	1992	1995	1998	2001	2003	2005	1992	1995	1998	2001	2003	2005
Average income at 2000 prices[5]												
Mean income (Euro)	23 961	24 180	24 803	24 998	24 966	24 918	15 877	15 765	16 118	16 519	16 398	16 871
Median income (Euro)	20 047	20 055	20 139	19 906	19 650	19 067	13 748	13 581	13 731	14 219	14 195	13 970
Relative difference[6] (%)	17.8	18.7	20.8	22.8	23.9	26.8	14.4	14.9	16.0	15.0	14.4	18.9
	Summary measures of inequality						Summary measures of inequality					
Gini coefficient[7]	0.3831	0.3838	0.3942	0.4016	0.4016	0.4115	0.3404	0.3401	0.3454	0.3465	0.3422	0.3631
Generalized entropy measures[7] [8]												
GE(0)	0.3264	0.3426	0.3603	0.3577	0.3580	0.3707	0.2745	0.2889	0.2949	0.2889	0.2831	0.3166
GE(1)	0.3122	0.3053	0.3307	0.3446	0.3357	0.3770	0.2561	0.2626	0.2798	0.2801	0.2717	0.3240
GE(2)	1.8552	2.2965	3.1315	3.4213	7.1995	10.4372	1.3182	2.3984	2.5392	3.2625	8.2937	13.0851
	Structure by income fractiles[9] in percent						Structure by income fractiles[9] in percent					
1st - 5th decile	30.81	30.33	29.97	29.68	29.86	30.38	34.86	34.95	35.43	35.18	35.65	35.81
6th - 9th decile	45.37	45.91	45.48	45.09	45.13	43.65	42.98	42.83	42.26	42.35	42.43	41.20
10th decile	23.82	23.76	24.55	25.22	25.01	25.97	22.16	22.22	22.31	22.47	21.91	22.99
Top 1%	6.25	5.85	6.31	6.63	6.16	7.08	5.76	5.73	6.05	6.15	5.77	6.76
Top 0.1%	2.30	2.11	2.38	2.54	2.26	2.86	1.95	1.99	2.21	2.32	2.14	2.77
Top 0.01%	0.89	0.86	1.01	1.07	1.00	1.31	0.73	0.80	0.92	0.98	0.97	1.30
Top 0.001%	0.30	0.32	0.40	0.42	0.46	0.58	0.24	0.30	0.36	0.40	0.46	0.58
Top 0.0001%	0.08	0.11	0.13	0.13	0.20	0.24	0.07	0.10	0.11	0.12	0.20	0.25
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

1) Income from business activity, wage income, capital income, exclusive public and private pensions; measured at the individual level.- 2) Gross market income plus transfer income.- 3) Married couples: half of the joint income is assigned to each spouse.- 4) Gross income less social security contributions, income tax and solidarity surcharge.- 5) Deflated by consumer price index.- 6) Difference of ln(mean) and ln(median).- 7) In cases with zero or negative income this income is replaced by 1 Euro.- 8) GE(0) is the mean logarithmic deviation, GE(1) is the Theil index, and GE(2) is half the square of the coefficient of variation.- 9) Ranking according to gross marked income, separated income of spouses.

Source: Stefan Bach, Giacomo Corneo, Viktor Steiner (2009): From Bottom to Top: The Entire Income Distribution in Germany, 1992 - 2003. Review of Income and Wealth 55 (2), 331-359.

Average income tax rates, 1992-2007

Gross income[1] plus local business tax liability fractiles	Assessed income tax liability (including solidarity surcharge) plus local business tax liability													
	in percent of gross income[1] plus local business tax liability less deducted losses carried forward/back							in percent of taxable income[2] plus local business tax liability						
	1992	1995	1998	2001	2004	2005	2007	1992	1995	1998	2001	2004	2005	2007
1st - 5th decile	3.7	3.4	2.4	2.1	1.6	1.7	1.8	10.8	11.7	9.2	8.0	6.9	7.7	4.9
6th - 9th decile	10.1	10.3	10.2	9.7	9.7	8.9	10.0	17.7	18.9	19.2	18.5	18.4	17.7	17.6
10th decile	22.8	21.1	22.8	23.4	22.1	21.9	24.2	31.9	31.8	34.0	33.7	31.8	31.6	33.0
Top 1%	38.2	34.3	34.8	37.3	34.3	34.0	35.5	46.4	46.6	44.6	45.2	42.3	40.9	41.7
Top 0.1%	46.6	42.8	40.4	44.3	39.7	38.9	39.7	53.0	53.9	47.2	49.8	46.4	44.2	44.9
Top 0.01%	48.4	45.0	41.4	45.8	39.9	38.9	39.3	52.9	53.7	46.1	50.7	47.1	44.7	45.1
Top 0.001%	47.4	43.5	47.1	45.4	37.1	36.7	37.1	51.7	52.2	52.0	51.7	47.4	44.6	44.9
Top 0.0001%	49.7	40.6	51.7	41.8	31.5	32.3	33.0	55.6	57.2	54.9	51.9	47.2	43.1	44.3
Total	12.8	12.3	12.8	12.7	12.3	11.9	14.2	22.4	23.1	24.4	24.0	23.2	23.1	23.6

1) For the definition of gross income, see Section 4. Top percentiles are nested within the preceding percentiles.- 2) Less child allowance.
Source: ITR-SOEP data base.

Source: Stefan Bach, Giacomo Corneo, Viktor Steiner (2013): Effective Taxation of Top Incomes in Germany. German Economic Review 14 (2), 115-137.

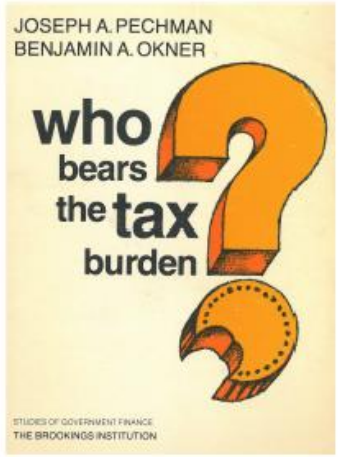
Who bears the tax burden in Germany?

Comprehensive analysis of the German tax system's distributional effects

- Including corporate taxation and indirect taxes
- Including very top income households, based on income tax files

Consistent microdata-based analysis

- Household surveys (SOEP, EVS), PIT files
- Data integration, updated to 2015
- Microsimulation models, incidence assumptions



Source: Stefan Bach, Martin Beznoska, Viktor Steiner: Who bears the tax burden in Germany? Tax structure slightly progressive. DIW Economic Bulletin 51+52.2016.

Integrated Data Base

SOEP and income and consumption survey (EVS)

Mahalanobis-Matching

Congruent matching variables

- Personal and household characteristics
- Special weight given to net household income

Personal income tax files

Weakly aggregated information on the top 10 percent of income distribution

Simulation of corporate income taxes based on distributed profits liable to personal income taxation

Integrated into SOEP data base, adjustment of weighting factors

Tax-Benefit Microsimulation Model (STSM+)

Personal income tax, corporate income taxes (on dist. profits)

Social security contributions

Social security transfers (in cash)

Consumption taxation modules

- VAT and insurance tax
- Excise taxes: energy taxes, taxes on tobacco, alcohol and gambling, motor vehicle tax, real estate taxes
- Excise taxes on production inputs

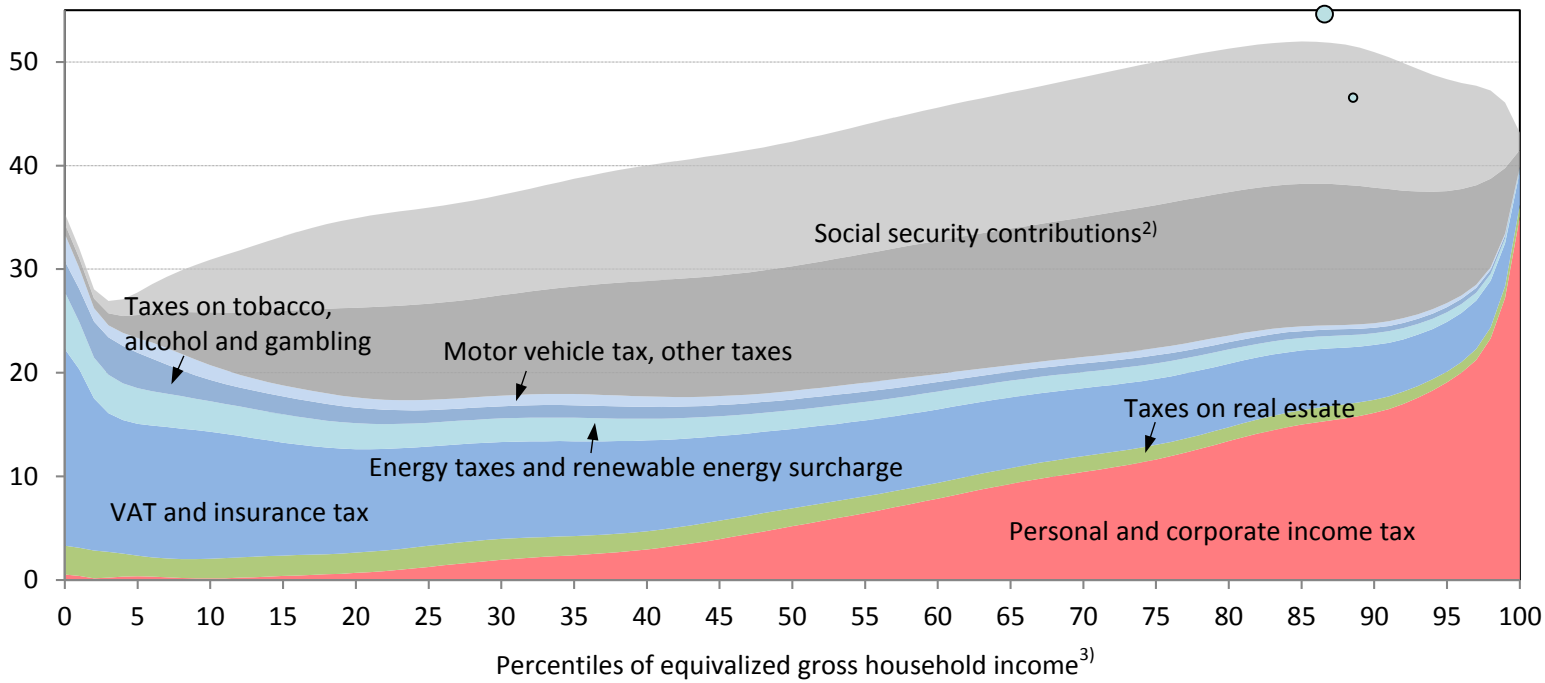
Behavioral models

- *AIDS based on income and consumption surveys*
- *Household labor supply model*

„Wale in a bathtub “ Wall Street Journal, Aug. 3, 2017.

Taxes and social security contributions as percent of gross household income, 2015[1)]

Integrated data base (SOEP, EVS, income tax statistics)

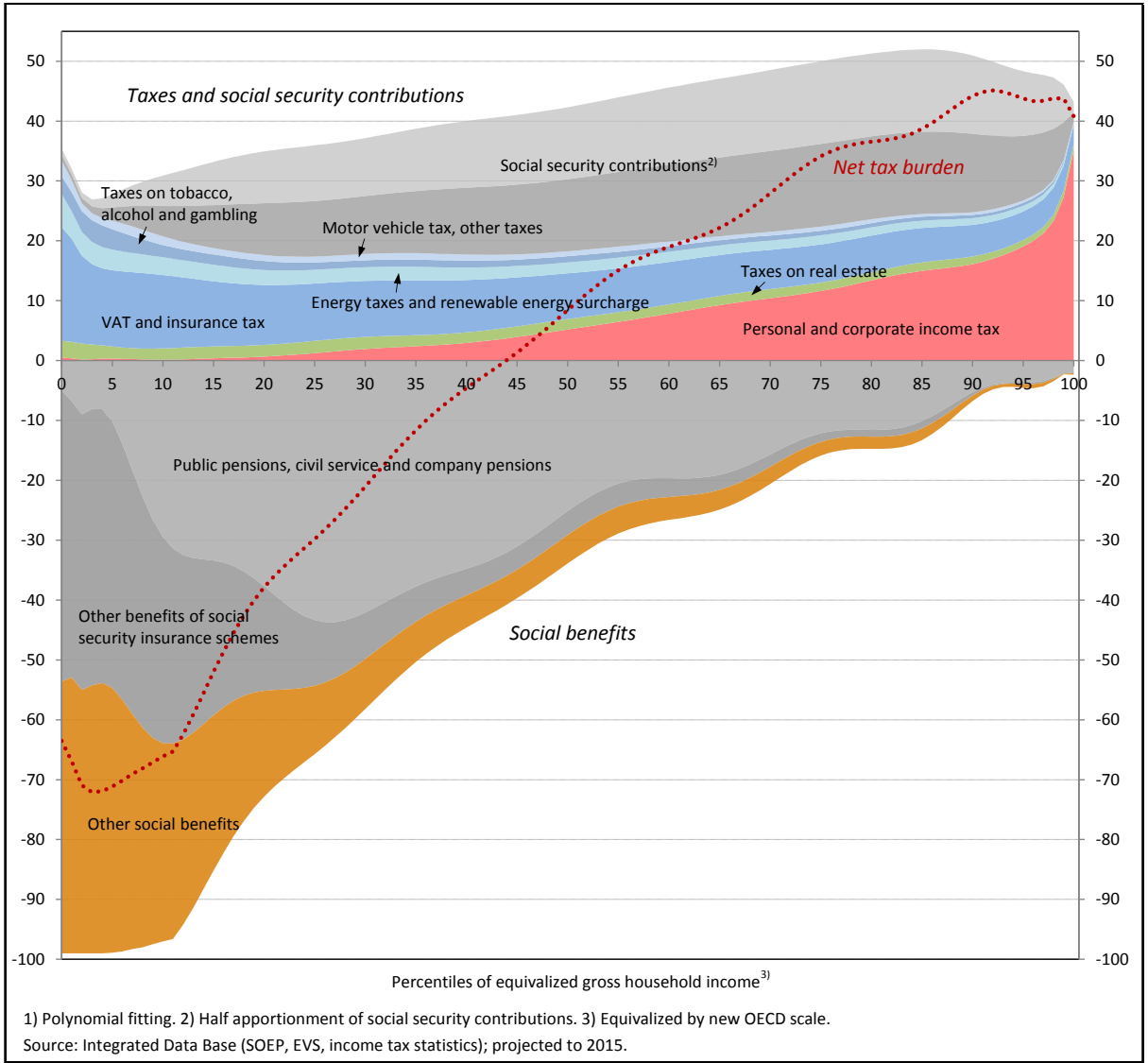


1) Polynomial fitting. 2) Half apportionment of social security contributions. 3) Equivalized by new OECD scale.

Source: Integrated Data Base (SOEP, EVS, income tax statistics); projected to 2015.

Source: Stefan Bach, Martin Beznoska, Viktor Steiner: Who bears the tax burden in Germany? Tax structure slightly progressive. DIW Economic Bulletin 51+52.2016.

Taxes, social security contributions, and social benefits as percent of gross household income, 2015[1)]
Integrated data base (SOEP, EVS, income tax statistics)



1) Polynomial fitting. 2) Half apportionment of social security contributions. 3) Equivalized by new OECD scale.
Source: Integrated Data Base (SOEP, EVS, income tax statistics); projected to 2015.

Prospect: DINA Germany – Distributional National Accounts

Personal income distribution consistent with national accounts

Covering the entire distribution from bottom to top

Including top incomes and (retained) corporate income

Secondary income: redistributive effect of the welfare state

Data

Income tax files, household surveys, data integration

Focus for Germany

Corporate income presumably strongly concentrated at the top, even more than in the U.S.

- High income concentration at the top
- “German Mittelstand”, “hidden champions”: closely held family firms
- High household savings, macroeconomic imbalances

Increasing income concentration at the top

Thank You for Your Attention!

sbach@diw.de
http://www.diw.de

@SBachTax