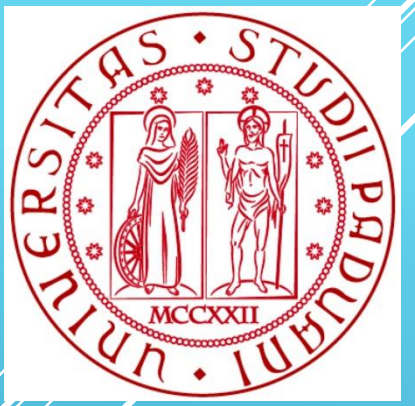


Sampling error and fixed effect estimation bias. A montecarlo simulation

Economic and Finance
Ch. mo. Prof. Luca Nunziata

Student
Clemente Cortile

Attenuation Bias

Main problem

Inadequate sampling rates on independent variables will cause bias in the estimaton of fixed effect models

Objective

Identify an optimal cell size per region to minimize bias

Test and compare the effectiveness of econometric models against the attenuation bias

Examples

- Contadictory results for the effect of regional immigration on wages level (Aydemir, Borjas 2012) due to cell size being too small
- Results of Nobel-prize winning study in physics (Perlmutter S. Et al, 1998) contested by a replication study perfomed on a bigger sample (Nielsen J.T., 2015)

Theoretical Framework

Attenuation Bias Structure (Aydemir, Borjas 2012)

For immigraton on wages FE model:

$$w_k = \beta\pi_k + \Sigma_h \alpha_h Z_{kh} + \varepsilon_k$$

True immigration:

$$p_k = \pi_k + u_k$$

$$\frac{plim\ \hat{\beta} - \beta}{\beta} = (1-\tau)\frac{\bar{p}\,(1-\bar{p}\,)/\bar{n}}{(1-R^2)\sigma_p^2}$$

τ	Sampling rate
$\bar{p}$	Average immigrant share
$\bar{n}$	Average cell size
$(1-R^2)\sigma_p^2 = \sigma_{res}^2$	Residual variance of immigrant share on wages

Key parameter

EMPIRICAL DATASET LIMITATION: only two population sample available (Canada & US data) to test the bias

Research design

Objective

Simulate a population model to avoid dataset's limitation and track the bias affecting estimated models

Simulation Structure - Population of an immigration on crime model (Nunziata, 2015)

Generate a population, redistribute it over regions, assigning immigrant status (average immigration share measure)

Simulate immigration shocks in form of waves to be added to the initial population across regions

Generate population FE model of immigration on crime by imposing it to match the data:

$$C_{rt} = m_{rt}\,\beta + \mu_r + \gamma_t + \varepsilon_t$$

$\overline{m}$ → σ_m → C

Population beta coefficient will be normalized to one to track the bias

DESIGN PORTABILITY: framework can be applied to any pair of econometric variable as long as they can be modeled as FE and there's a statistically significant relationship

Research design

Simulation Structure – Monte Carlo simulation on estimated model

Simulated Populations → MCS at different sampling rates → Estimated models

BASELINE FIXED EFFECT MODEL

$$C_{rt} = m_{rt}\beta_{FE} + \mu_r + \gamma_t + \varepsilon_t$$

- Can account for unobserved time invariant characteristic, but is subject to attenutation bias

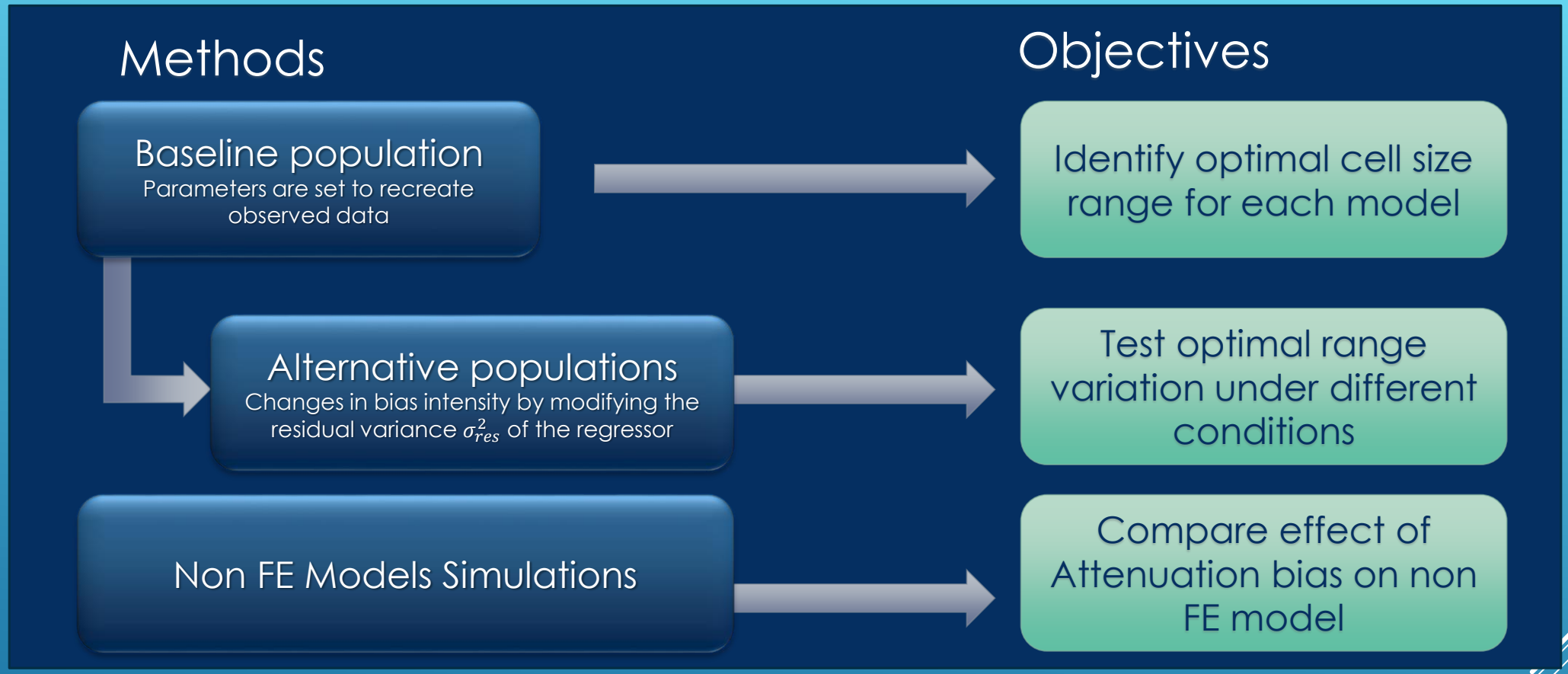
SPLIT SAMPLE INSTRUMENTAL VARIABLE MODEL

$$\begin{cases} C_{rt} = \widetilde{m_{rt}^{X}}\beta_{IV} + \mu_r + \gamma_t + \varepsilon_t \\ \widetilde{m_{rt}^{X}} = m_{rt}^{Z}\beta_{FSIV} + \eta_t \end{cases}$$

- Can account for unobserved time invariant characteristic, is less susceptible to attenutation bias, but requires an adequate instrument for the first stage estimation

$|1 - \beta_{FE}|$ and $|1 - \beta_{SSIV}|$ will be a measure of the bias for every estimated model.

Simulation Outline

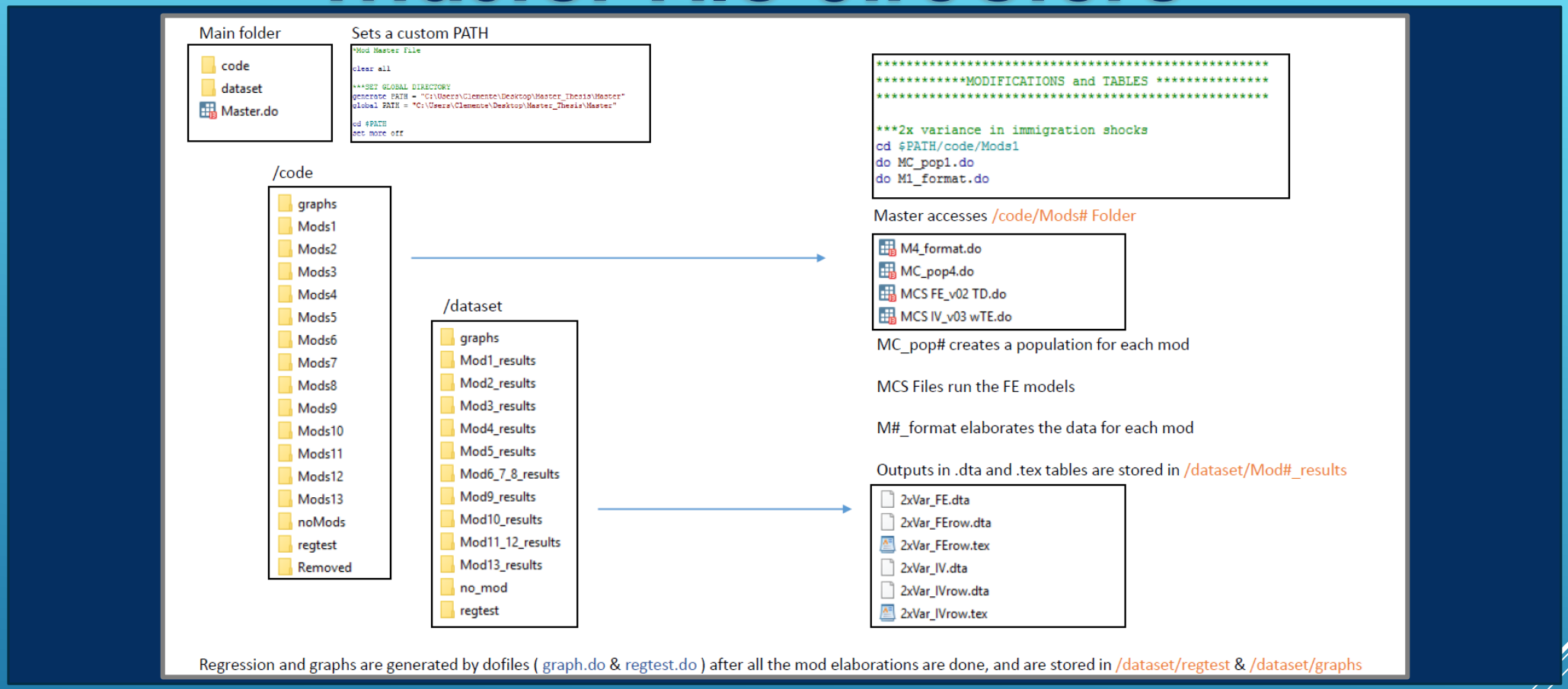


Simulation Statistics

Total simulation computing time	***2544 hrs (106 days)
Total number of individuals	321'477'100
Models simulated	2400
Amount of data generated	2.27 TB

****Runtimes on a desktop IntelCore ·i5-4430 Processor at 3.20 GHz - 16GB RAM - Win10 x64 by STATA MP14*

Master File Structure



Simulation Statistics

Total simulation computing time	***2544 hrs (106 days)
Total number of individuals	321'477'100
Models simulated	2400
Amount of data generated	2.27 TB

*****Runtimes on a desktop IntelCore ·i5-4430 Processor at 3.20 GHz - 16GB RAM - Win10 x64 by STATA MP14*

Baseline population

1) Baseline population parameters are set to recreate actual data conditions

BASELINE POPULATION SETUP	
Initial pop. count: 10'000'000	Regions: 100
Immigration Shocks intensity: 1% increase yearly *(w.r.t. Initial pop count)*	Average proportion of immigrants: 0,10
Time length: 4 *(8 year)*	Dependent variable error term distr. : $N\sim(0;0,5)$

2) Only parameters affecting the sampling rate are modified to find the optimal cell size

Key parameters	Parameter value	Cell range FE model	Cell range SSIV model First stage variable	Cell range SSIV model Second stage variable
Initial Population size	100'000 500'000; 1'000'000; 10'000'000;	(3, 209) (5, 522) (10, 1043) (10, 31299)	(3, 209) (5, 522) (10,1043) (104,10433)	(1, 14) (5, 52) (14, 104) (104, 1043)
Regions	50; 100	(209, 20866) (104, 10433)	(209, 20866) (104,10433)	(209, 280) (104, 1043)
SSIV Endogenous variable sampling rate (for immigration waves variance = 4%)			(104, 10433)	(10, 1020)

Baseline population - FE

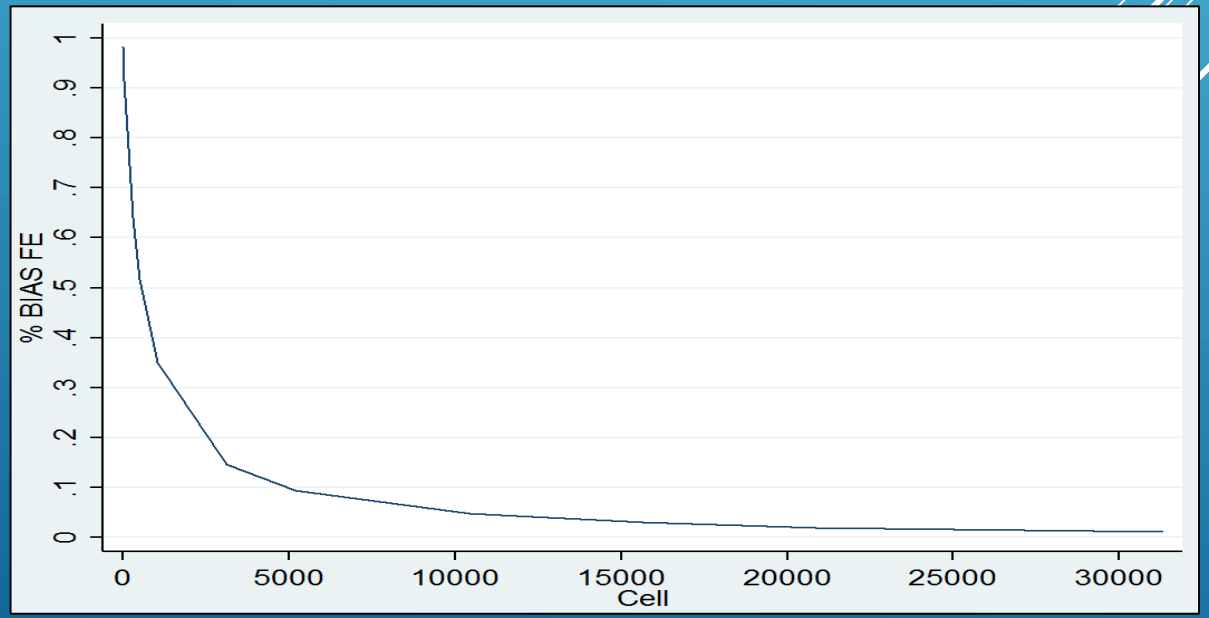
Table 1 – FE Coefficients

S.Rate	0.01	0.05	0.1	0.3	0.5	1	3	5	10	15	20	30
$\widehat{\beta}^{FE}_{popul}$	.997	.998	.996	1	1	1	1	1	1	1	1	1
s.e.	.0721	.0322	.0228	.0131	.0102	.00719	.00415	.00322	.00228	.00186	.00161	.00131
$\widehat{\beta}^{FE}_{sample}$	.0184	.0862	.153	.36	.483	.65	.854	.905	.952	.97	.981	.989
s.e.	.00995	.00955	.00916	.008	.00713	.00585	.00386	.00306	.00222	.00183	.00159	.00131
Obs Cell	10.4	52.2	104	313	522	1043	3130	5216	10433	15649	20866	31299
s.e.	.53	2.14	4.28	12.8	21.4	42.8	128	214	428	642	856	1283

Cell size	100	300	500	800	1000	2000	3000	5000
Bias w.r.t. population coefficient	89%	76%	63%	48%	42%	21%	14%	10%

- FE is affected by a bias bigger than 50% for cell sizes smaller than 500: *the effect of immigration on crime rate is halved*
- Above 5000 obs. per cell the bias is negligible

Graph 1 – % Bias vs Avg cell size – Baseline FE



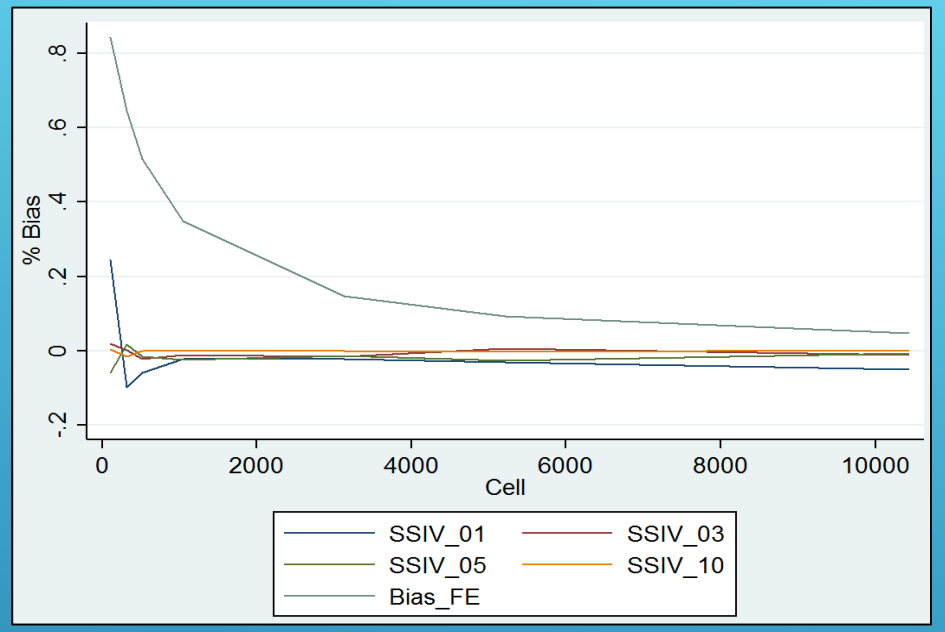
Baseline population: SSIV vs FE

Table 2 – FE vs SSIV Coefficients

S.Rate	0.1	0.3	0.5	1	3	5	10
$\widehat{\beta}^{FE}_{popul}$	1	.998	1	.999	.999	1	1
s.e.	.0198	.0114	.00886	.00626	.00362	.0028	.00198
$\widehat{\beta}^{FE}_{sample}$	.156	.354	.484	.651	.852	.908	.953
s.e.	.00801	.00693	.00628	.0051	.00335	.00268	.00194
$\widehat{\beta}^{IV}_{sample\ 0.1}$	.755	1.1	1.06	1.02	1.02	1.03	1.05
s.e.	.294	.0257	.0163	.00932	.00468	.00356	.0025
$\widehat{\beta}^{IVFS}_{sample\ 0.1}$	.154	.341	.482	.658	.849	.901	.936
s.e.	.00479	.00416	.00374	.00299	.00196	.00154	.0011
$\widehat{\beta}^{IV}_{sample\ 0.3}$	.981	.997	1.02	1.01	1.02	.993	1.01
s.e.	.0527	.0203	.0138	.00829	.0042	.00309	.00217
$\widehat{\beta}^{IVFS}_{sample\ 0.5}$	.166	.361	.48	.65	.847	.921	.949
s.e.	.00313	.00261	.00232	.00182	.00115	.000914	.000648
$\widehat{\beta}^{IV}_{sample\ 0.5}$	1.06	.984	1.02	1.02	1.01	1.03	1.01
s.e.	.0553	.0196	.0134	.00821	.00411	.00313	.00211
$\widehat{\beta}^{IVFS}_{sample\ 0.5}$	.151	.364	.481	.641	.844	.889	.95
s.e.	.00266	.0022	.00193	.00149	.000915	.000712	.000503
$\widehat{\beta}^{IV}_{sample\ 1}$	.997	1.01	1	.999	1	1	.998
s.e.	.0513	.0198	.013	.00788	.00399	.003	.00206
$\widehat{\beta}^{IVFS}_{sample\ 1}$	.158	.351	.485	.654	.853	.907	.957
s.e.	.00225	.00182	.00157	.00116	.000673	.000521	.00036
Obs Cell	104	313	522	1043	3130	5216	10433

- SSIV *is mostly unaffected*: bias for cell sizes sizes bigger than 100 is smaller than 1%
- Raising the instrument variable cell size also affects the coefficient for the first stage

Graph 2 – % Bias vs Avg cell size – Baseline SSIV



Optimal range criteria

- *Reduction in bias by 90 percentage point*
- *Higher cell size ranges have the same or lower percentage of bias (consistent) and are below 10 percentage points*
- *Estimates are diagonally mirrored within a 10 percentage points (SSIV only)*

SSIV: instrument swap

Table 5 – Instrument swap - SSIV

S.Rate	0.1	0.3	0.5	1	3	5	10
$\widehat{\beta}^{FE}_{popul}$	1	.998	1	.999	.999	1	1
s.e.	.0198	.0114	.00886	.00626	.00362	.0028	.00198
$\widehat{\beta}^{FE}_{sample}$	.155	.361	.487	.654	.855	.914	.954
s.e.	.00806	.00697	.00632	.00511	.00336	.0027	.00194
$\widehat{\beta}^{IV}_{sample\ 0.1}$	1.27	.988	1.06	.964	1.01	1.01	1
s.e.	.134	.0309	.025	.0156	.00941	.00753	.00516
$\widehat{\beta}^{IVFS}_{sample\ 0.1}$	.16	.162	.154	.166	.155	.151	.155
s.e.	.00485	.00183	.0012	.000721	.000355	.000266	.000184
$\widehat{\beta}^{IV}_{sample\ 0.3}$	1.12	1.05	1.03	.98	1	1	.998
s.e.	.047	.0212	.0159	.0105	.00607	.00476	.0033
$\widehat{\beta}^{IVFS}_{sample\ 0.3}$	.347	.35	.351	.366	.361	.354	.362
s.e.	.00723	.00263	.0017	.001	.000481	.000353	.000243
$\widehat{\beta}^{IV}_{sample\ 0.5}$	1.04	.995	1.02	.997	1.01	1	.998
s.e.	.0354	.0173	.0135	.00916	.00528	.00405	.00282
$\widehat{\beta}^{IVFS}_{sample\ 0.5}$	.487	.493	.484	.485	.477	.485	.49
s.e.	.00828	.00299	.00193	.0011	.000513	.000374	.000253
$\widehat{\beta}^{IV}_{sample\ 1}$	1.03	.995	1.01	.992	1	1.01	.999
s.e.	.0295	.0149	.0115	.00781	.0045	.00354	.00244
$\widehat{\beta}^{IVFS}_{sample\ 1}$	.648	.657	.645	.66	.652	.641	.656
s.e.	.00939	.00332	.00208	.00117	.000515	.000375	.000246
Obs Cell	104	313	522	1043	3130	5216	10433

Table 2 – FE vs SSIV Coefficients

S.Rate	0.1	0.3	0.5	1	3	5	10
$\widehat{\beta}^{FE}_{popul}$	1	.998	1	.999	.999	1	1
s.e.	.0198	.0114	.00886	.00626	.00362	.0028	.00198
$\widehat{\beta}^{FE}_{sample}$	.156	.354	.484	.651	.852	.908	.953
s.e.	.00801	.00693	.00628	.0051	.00335	.00268	.00194
$\widehat{\beta}^{IV}_{sample\ 0.1}$	.755	1.1	1.06	1.02	1.02	1.03	1.05
s.e.	.294	.0257	.0163	.00932	.00468	.00356	.0025
$\widehat{\beta}^{IVFS}_{sample\ 0.1}$	.154	.341	.482	.658	.849	.901	.936
s.e.	.00479	.00416	.00374	.00299	.00196	.00154	.0011
$\widehat{\beta}^{IV}_{sample\ 0.3}$	.981	.997	1.02	1.01	1.02	.993	1.01
s.e.	.0527	.0203	.0138	.00829	.0042	.00309	.00217
$\widehat{\beta}^{IVFS}_{sample\ 0.5}$	.166	.361	.48	.65	.847	.921	.949
s.e.	.00313	.00261	.00232	.00182	.00115	.000914	.000648
$\widehat{\beta}^{IV}_{sample\ 0.5}$	1.06	.984	1.02	1.02	1.01	1.03	1.01
s.e.	.0553	.0196	.0134	.00821	.00411	.00313	.00211
$\widehat{\beta}^{IVFS}_{sample\ 0.5}$	.151	.364	.481	.641	.844	.889	.95
s.e.	.00266	.0022	.00193	.00149	.000915	.000712	.000503
$\widehat{\beta}^{IV}_{sample\ 1}$	.997	1.01	1	.999	1	1	.998
s.e.	.0513	.0198	.013	.00788	.00399	.003	.00206
$\widehat{\beta}^{IVFS}_{sample\ 1}$	.158	.351	.485	.654	.853	.907	.957
s.e.	.00225	.00182	.00157	.00116	.000673	.000521	.00036
Obs Cell	104	313	522	1043	3130	5216	10433

- Swapping endogenous variable and instrument has no relevant effect as expected from the properties of SSIV, *however lower cell sizes have small differences*
- The biases are symmetric as expected

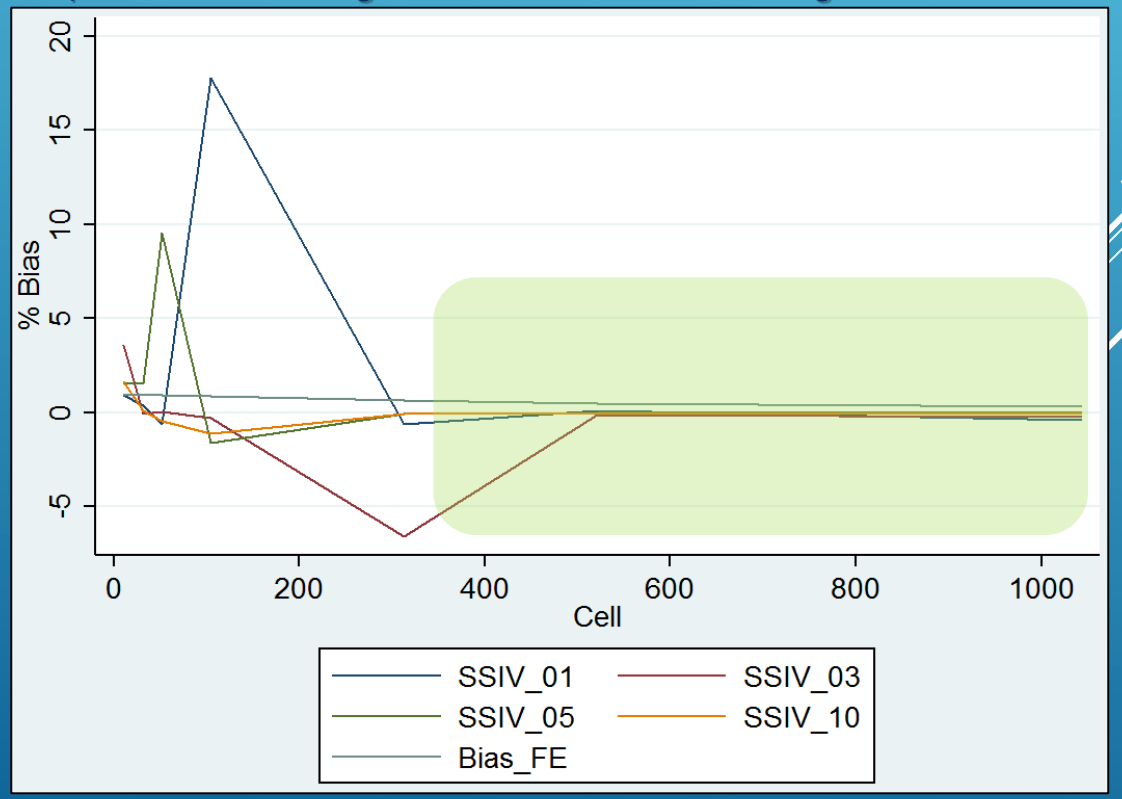
Baseline population – SSIV

Table 3 – Baseline SSIV – lower cell range

S.Rate	0.1	0.3	0.5	1	3	5	10
$\widehat{\beta}^{FE}_{popul}$	1.02	.997	.999	1	.996	.999	1
s.e.	.0629	.0362	.028	.0198	.0114	.00887	.00627
$\widehat{\beta}^{FE}_{sample}$	.0161	.0538	.0864	.156	.361	.495	.677
s.e.	.00876	.00851	.00833	.00816	.00711	.00635	.00523
$\widehat{\beta}^{IV}_{sample\ 0.1}$	.0737	.634	1.63	16.8	1.62	.89	1.39
s.e.	.832	7.96	8.23	2465	.251	.182	.0681
$\widehat{\beta}^{IVFS}_{sample\ 0.1}$	.0122	.0297	.0937	.137	.366	.499	.647
s.e.	.0158	.0152	.0149	.0145	.0129	.0115	.00936
$\widehat{\beta}^{IV}_{sample\ 0.3}$	-2.58	1.07	.948	1.27	7.59	1.14	1.22
s.e.	106	2.13	2.44	.262	32.6	.0309	.0178
$\widehat{\beta}^{IVFS}_{sample\ 0.3}$	.0164	.0481	.0783	.163	.328	.531	.638
s.e.	.0092	.00883	.00866	.0085	.00736	.00663	.00546
$\widehat{\beta}^{IV}_{sample\ 0.5}$	-.524	-.547	-8.52	2.62	1.04	.986	1.03
s.e.	14	4.76	513	3.78	.0281	.0163	.0106
$\widehat{\beta}^{IVFS}_{sample\ 0.5}$	.0239	.0473	.0739	.154	.376	.518	.702
s.e.	.00722	.00698	.00681	.00673	.00583	.00514	.0042
$\widehat{\beta}^{IV}_{sample\ 1}$	-.665	.893	1.46	2.1	1.07	1.06	1.08
s.e.	9.03	.585	.537	1.71	.0253	.016	.00986
$\widehat{\beta}^{IVFS}_{sample\ 1}$	.0157	.0463	.0794	.154	.361	.489	.648
s.e.	.00533	.00513	.005	.0049	.00418	.00375	.00306
Obs Cell	10.4	31.3	52.2	104	313	522	1043

- Lower ranges for SSIV *are heavily affected* by bias for cell sizes sizes lower than 50 for the endogenous variable and 300 for the instrument
- Lower cell sizes for the two stages cause biased coefficient to randomly arise across the table

Graph 3 – % Bias vs Avg cell size – SSIV lower cell range



Baseline population – SSIV 2

Table 4 – Optimal cell size by range - SSIV

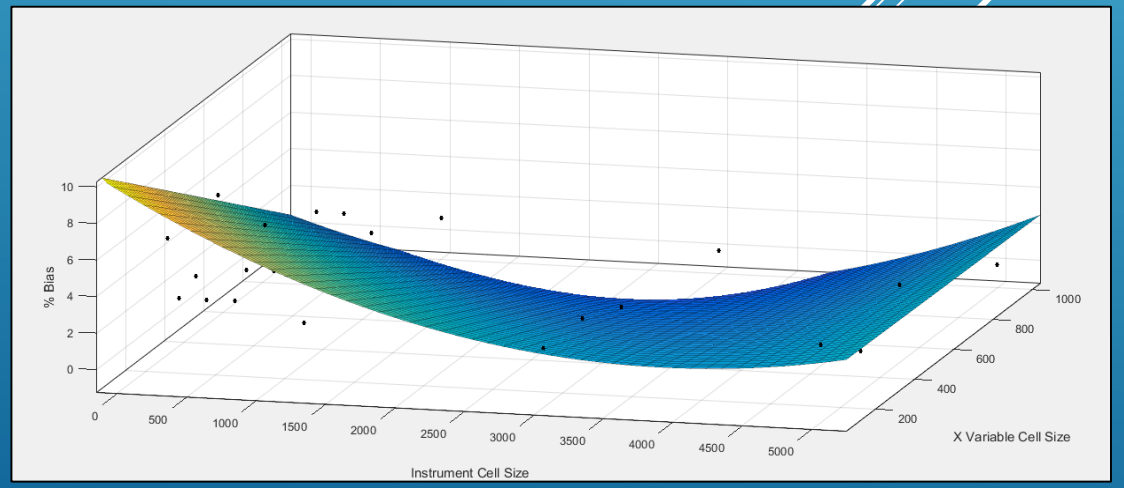
Second stage regressor	First Stage regressor								
Cell size	10	30	50	100	150	200	300	500	1000
10	92	36	- 63	177	86	7	-62	10	-39
30	358	-7	5	-26	30	32	28	-13	-22
50	152	154	952	-162	-17	1	-4	1	-2
100	166	10	-45	25*	25	1	-7	-5	-8
150	45	95	1	13	17	10	2	-3	-3
200	40	-659	4	7	1	-1	1	-2	-2
300	3	35	-3	2	-2	2	2	-2	-1
500	100	28	-3	6	-3	1	2	-2	-1
1000	150	1	6	1	1	1	1	0	1

**Coefficient has a 25% bias with a t-stat > 2 in simulation ESS*

- Green area marks optimal cell sizes
- The orange area doesn't show a consistent reduction in bias
- Light green area is the lower bound for optimal cell sizes

$$f(x,y) = 10{,}47 - 0.04292x - 9{,}996\,10^{-2}y + 5{,}34\,10^{-7}x^2 - 1{,}18\,10^{-6}x\,y$$

Goodness of fit: SSE: 369.4; R-square: 0.3827



2.2 Modified population model

Population list by parameters and average cell sizes covered in the simulation

Key parameters	Parameter value	Cell range FE model	Cell range SSIV model Z variable	Cell range SSIV model X variable
Initial average immigrant share (as % of total population)	0.05% 10%; 30%; 50%;	(10, 6129) (10, 31299) (113, 5649) (122, 6082)	(102, 10200) (104, 10433) (113, 5649) (122, 6082)	(102, 2043) (104, 1043) (113,1130) (122, 1216)
Immigration shock intensity (as % of yearly increase in population)	0.5%; 1%; 2.5%; 6.75%;	(10, 30599) (10, 31299) (11, 33062) (127, 12679)	(102, 10200) (104, 10433) (110, 11009) (127, 12679)	(102, 1020) (104, 1043) (110, 1101) (127,1268)
Panel data lenght (unit of time = 2 years)	2 4 8	(10, 20472) (10, 31299) (15, 30355)	(102, 10236) (104, 10433) (152, 15178)	(102, 1024) (104, 1043) (152, 1518)
SSIV Endogenous variable sampling rate (for immigration waves variance = 10%)			(104, 10433)	(10, 1043)
Smaller variance in the error term for dependent variable		(10, 31299)	(104, 10433)	(104, 1043)

Baseline parameters value

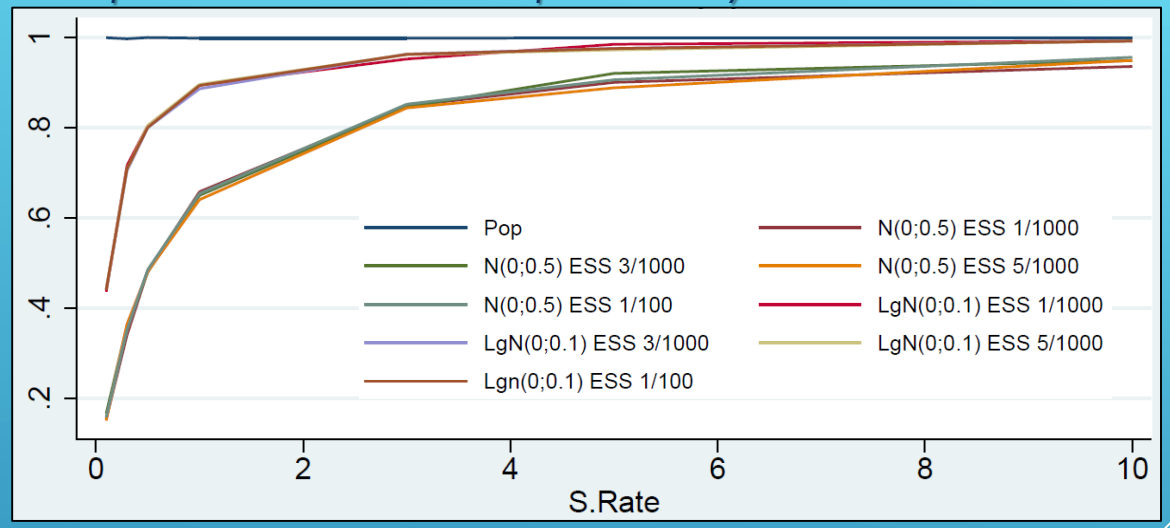
DESIGN PORTABILITY: framework can be applied to any pair of econometric variable as long as they can be modeled as FE and there's a statistically significant relationship

SSIV: reduced error term

Table 7 – Reduced error term - SSIV

S.Rate	0.1	0.3	0.5	1	3	5	10
$\widehat{\beta}^{FE}_{popul}$	.999	1	1	1	1	1	1
s.e.	.00879	.00508	.00393	.00278	.00161	.00124	.00088
$\widehat{\beta}^{FE}_{sample}$	.439	.706	.804	.892	.962	.976	.989
s.e.	.00633	.00447	.00363	.00267	.00159	.00123	.000877
$\widehat{\beta}^{IV}_{sample\ 0.1}$	1.02	.993	1.01	1	1.02	.995	1
s.e.	.0161	.00706	.00519	.00347	.00196	.00147	.00104
$\widehat{\beta}^{IVFS}_{sample\ 0.1}$	.436	.717	.805	.895	.953	.986	.993
s.e.	.00426	.0029	.0023	.00167	.000969	.000755	.000535
$\widehat{\beta}^{IV}_{sample\ 0.3}$	1	1	1.01	1.01	1	1	.998
s.e.	.0142	.00646	.00467	.00315	.00174	.00134	.000936
$\widehat{\beta}^{IVFS}_{sample\ 0.3}$	.442	.705	.801	.887	.964	.976	.993
s.e.	.00314	.00196	.00149	.00104	.000583	.000444	.00031
$\widehat{\beta}^{IV}_{sample\ 0.5}$	.994	.998	1	.996	1	1	.997
s.e.	.0138	.00627	.00455	.00304	.0017	.00131	.000915
$\widehat{\beta}^{IVFS}_{sample\ 0.5}$	.442	.708	.805	.897	.963	.973	.993
s.e.	.00287	.00172	.00128	.000852	.000461	.000352	.000244
$\widehat{\beta}^{IV}_{sample\ 1}$	.997	.996	1	.998	.999	1	.997
s.e.	.0136	.00615	.00449	.003	.00167	.00128	.000898
$\widehat{\beta}^{IVFS}_{sample\ 1}$	.441	.71	.801	.894	.963	.976	.993
s.e.	.00265	.00152	.00109	.000691	.000351	.000255	.000175
Obs Cell	110	330	551	1101	3303	5505	11009
s.e.	10.9	32.7	54.4	109	327	545	1089
N.of Obs.	400	400	400	400	400	400	400

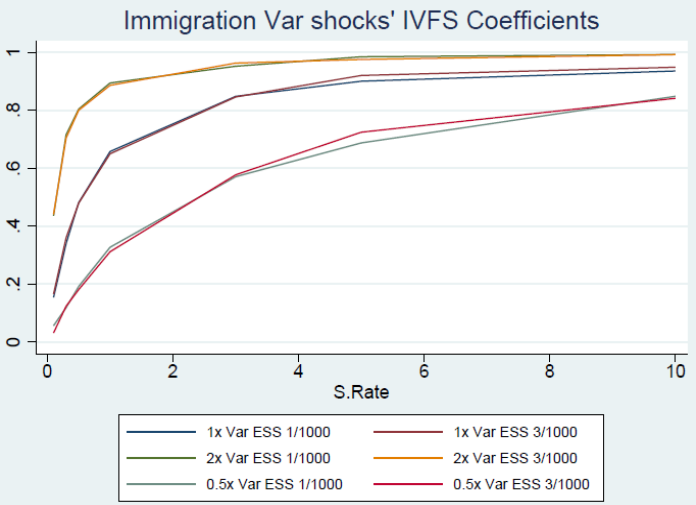
Graph – SSIV Betas comparison by error terms

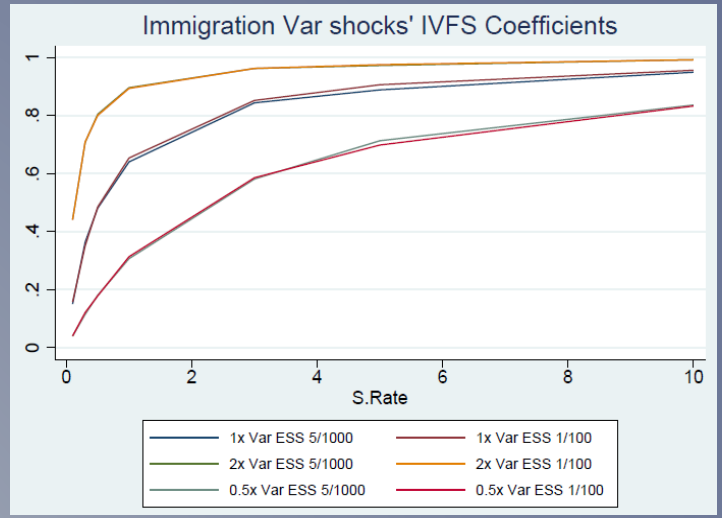


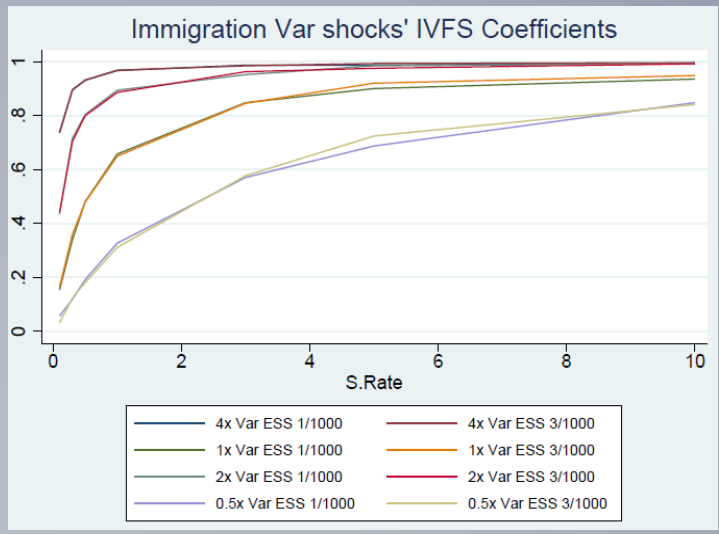
- Reducing the error term in the population model, has the same effect as changing the explanatory power of the regressor

SSIV: first stage quality

Graph 1 – % Bias vs Avg cell size – Baseline FE







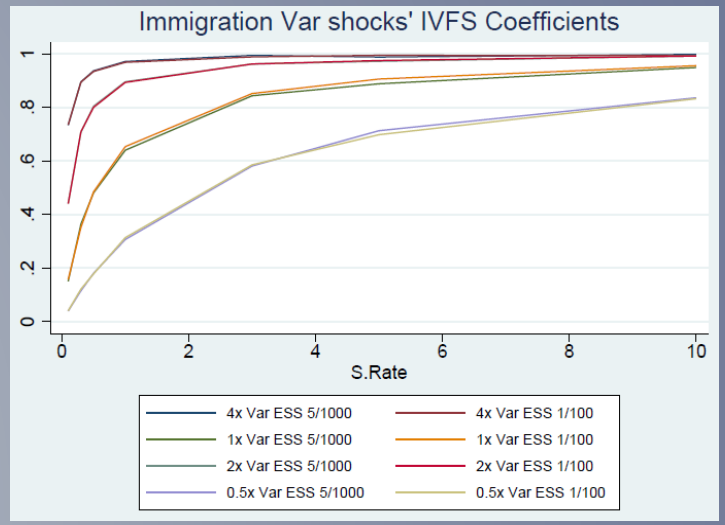


Table 6 – Instrument quality - SSIV

S.Rate	0.1	0.3	0.5	1	3	5	10
0.5xVar FSIV ESS 1/1000	.0566	.119	.192	.328	.571	.688	.849
0.5xVar FSIV ESS 3/1000	.0317	.124	.181	.312	.578	.725	.842
0.5xVar FSIV ESS 5/1000	.0398	.115	.181	.307	.582	.714	.837
0.5xVar FSIV ESS 1/100	.0407	.121	.179	.314	.586	.699	.833
1xVar FSIV ESS 1/1000	.154	.341	.482	.658	.849	.901	.936
1xVar FSIV ESS 3/1000	.166	.361	.48	.65	.847	.921	.949
1xVar FSIV ESS 5/1000	.151	.364	.481	.641	.844	.889	.95
1xVar FSIV ESS 1/100	.158	.351	.485	.654	.853	.907	.957
2xVar FSIV ESS 1/1000	.436	.717	.805	.895	.953	.986	.993
2xVar FSIV ESS 3/1000	.442	.705	.801	.887	.964	.976	.993
2xVar FSIV ESS 5/1000	.442	.708	.805	.897	.963	.973	.993
2xVar FSIV ESS 1/100	.441	.71	.801	.894	.963	.976	.993
4xVar FSIV ESS 1/1000	.739	.897	.932	.968	.988	.988	.996
4xVar FSIV ESS 3/1000	.736	.896	.933	.969	.986	.994	.997
4xVar FSIV ESS 5/1000	.734	.896	.938	.972	.995	.989	.999
4xVar FSIV ESS 1/100	.736	.895	.935	.969	.989	.995	.994

FSIV quality factors:

- increase in explained variance of the model
- bigger cell size for the instrument

2.2 Modified population: FE (init.avg)

Exponential fit for modified population model: average prop. of immigrants

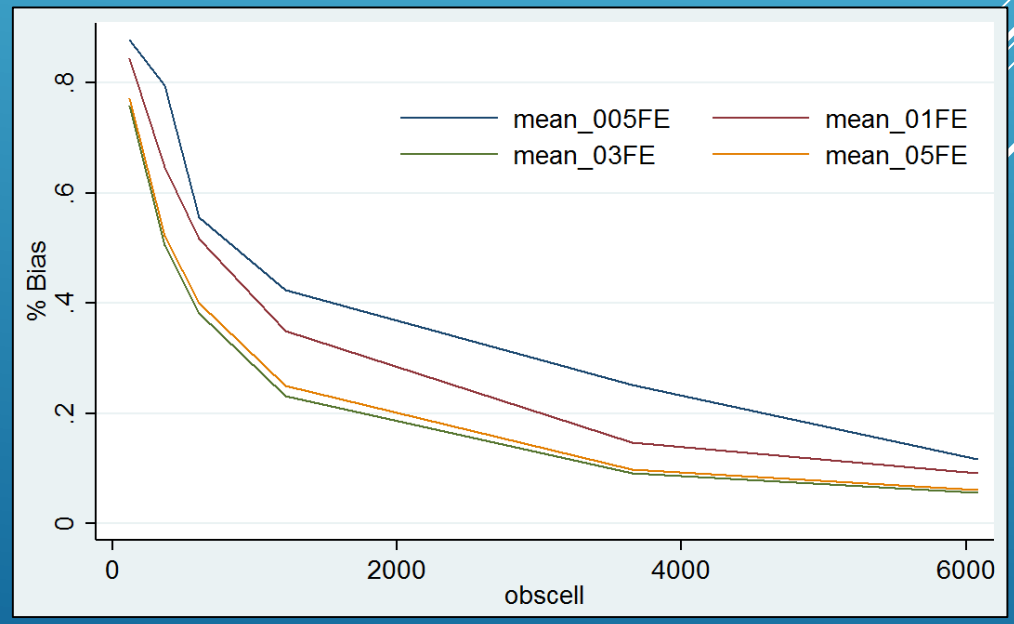
0.05	0.1	0.3	0.5
Coefficients (with 95% confidence bounds): a = 0.6882 (0.5572, 0.8192) b = -0.001337 (-0.001753, -0.000921) c = 0.2805 (0.1347, 0.4263) d = -0.000144 (-0.0002496, -3.848e-05) Goodness of fit: SSE: 8.307e-05 R-square: 0.9998	Coefficients (with 95% confidence bounds): a = 0.6187 (0.4045, 0.833) b = -0.002168 (-0.003423, -0.0009124) c = 0.3495 (0.1042, 0.5947) d = -0.0002865 (-0.0004953, -7.782e-05) Goodness of fit: SSE: 0.0001905 R-square: 0.9996	Coefficients (with 95% confidence bounds): a = 0.6513 (0.5016, 0.8009) b = -0.002686 (-0.004088, -0.001283) c = 0.2978 (0.1229, 0.4727) d = -0.0003021 (-0.0004957, -0.0001085) Goodness of fit: SSE: 0.0002219 R-square: 0.9994	Coefficients (with 95% confidence bounds): a = 0.6349 (0.4697, 0.8001) b = -0.002667 (-0.004236, -0.001097) c = 0.3226 (0.1293, 0.5159) d = -0.0003015 (-0.0004981, -0.000105) Goodness of fit: SSE: 0.0002668 R-square: 0.9993
$f(x) = 0.6882e^{-0.001337x} + 0.2805e^{-0.000144x}$	$f(x) = 0.6187e^{-0.002168x} + 0.3495e^{-0.0002865x}$	$f(x) = 0.6513e^{-0.002686x} + 0.2978e^{-0.0003021x}$	$f(x) = 0.6349e^{-0.002667x} + 0.3226e^{-0.0003015x}$

Table 5: Bias % w.r.t. increase in average prop. of immigrants

Modified Populations	Cell size						
Cell size	100	300	500	1000	3000	5000	Overall Improvement (in % w.r.t baseline bias)
5%	87	72	61	42	19	13 (6650)*	-33%
10%	84	64	51	34	14	9 (4500)*	-
30%	77	52	39	24	12 (3600)*	5	+ 46%
50%	75	50	38	23	13 (3500)*	6	+ 44%

**Cell size for which estimated attenuation bias is below 90%*

Graph – % Bias vs cell size across modified populations



2.2 Modified population model - SSIV (init. avg)

Table 7 – Optimal cell size by range – 5% init.avg

Second stage regressor: Cell size \ First Stage regressor	50	100	300	500	1000	3000	5000
50	90	-73	-280	-39	-22	-3	-2
100	60	72	-41	-16	6	-3	-4
300	-99	91	-24	0	3	0	-1
500	-25	-31	-9	-7	0	0	3
1000	-22	-14	-8	-3	0	1	-1

Improvement w.r.t. baseline bias

- 1037%

+77%

MAX = 100%
MIN = - ∞

Table 8 – Optimal cell size by range – 30% init.avg

Second stage regressor: Cell size \ First Stage regressor	50	100	300	500	1000	3000	5000
50	48	-8	1	-7	-1	1	-1
100	-9	-11	-3	-2	-4	1	-2
300	1	-2	-1	-2	1	1	1
500	-6	0	-1	-1	1	1	0
1000	1	-1	0	1	-1	0	1

$$f(x,y) = \alpha + \gamma_1 X + \gamma_2 Y + \gamma_3 X^2 + \gamma_4 XY$$

Coefficients (with 95% confidence bounds):

α = 11.68 (-5.926, 29.29)
γ_1 = -24.01 (-43.81, -4.206)
γ_2 = -9.788 (-19.78, 0.2069)
γ_3 = 9.68 (-5.345, 24.71)
γ_4 = 5.163 (-4.979, 15.3)

Goodness of fit: R-square: 0.7403

Baseline – 10% init.avg

Second stage regressor: Cell size \ First Stage regressor	50	100	300	500	1000	3000	5000
50	952	-162	-4	1	-4	4	6
100	-45	25	-7	-5	-8	2	3
300	-3	2	2	-2	-1	2	1
500	-3	6	2	-2	-1	1	3
1000	6	1	1	0	1	0	0

$$f(x,y) = \alpha + \gamma_1 X + \gamma_2 Y + \gamma_3 X^2 + \gamma_4 XY$$

Coefficients (with 95% confidence bounds):

α = 1.267 (-1.852, 4.387)
γ_1 = -0.4953 (-4.003, 3.012)
γ_2 = -0.7163 (-2.487, 1.054)
γ_3 = 0.1794 (-2.482, 2.841)
γ_4 = 0.3535 (-1.443, 2.15)

Goodness of fit: R-square: 0.6659

2.2 Modified population: FE (shock int.)

Exponential fit for modified population model by immigration shock intensity

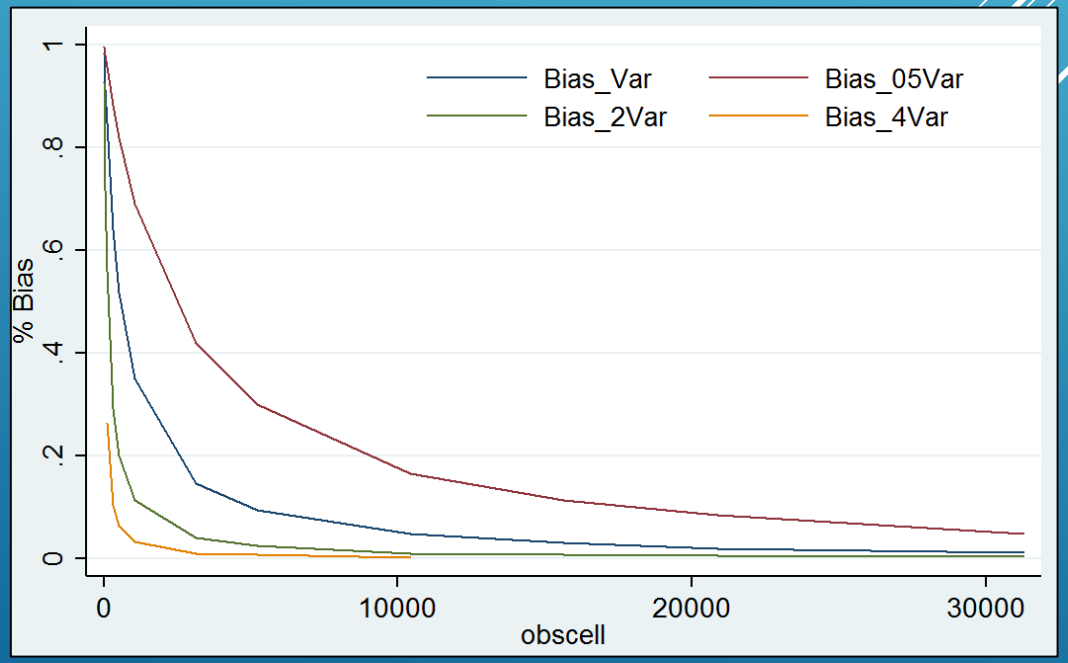
0.5%	1% (Baseline)	2.5%	6.75%
Coefficients (with 95% confidence bounds): a = 0.6026 (0.5382, 0.667) b = -0.0005706 (-0.0006632, -0.000478) c = 0.3923 (0.3254, 0.4592) d = -7.591e-05 (-8.963e-05, -6.22e-05) Goodness of fit: SSE: 0.0005837 R-square: 0.9996	Coefficients (with 95% confidence bounds): a = 0.7246 (0.6359, 0.8133) b = -0.001809 (-0.002171, -0.001448) c = 0.258 (0.1666, 0.3495) d = -0.0001686 (-0.0002434, -9.375e-05) Goodness of fit: SSE: 0.001652 R-square: 0.999	Coefficients (with 95% confidence bounds): a = 0.7056 (0.6344, 0.7768) b = -0.00787 (-0.009452, -0.006287) c = 0.2702 (0.197, 0.3434) d = -0.000691 (-0.0009964, -0.0003857) Goodness of fit: SSE: 0.001338 R-square: 0.9988	Coefficients (with 95% confidence bounds): a = 0.4022 (0.3426, 0.4618) b = -0.006308 (-0.008415, -0.004201) c = 0.05879 (0.02167, 0.09591) d = -0.0005108 (-0.0009509, -7.085e-05) Goodness of fit: SSE: 5.231e-05 R-square: 0.999
$f(x) = 0.6026e^{-0.0005706x} + 0.3923e^{-0.0000075x}$	$f(x) = 0.7246e^{-0.001809x} + 0.2580e^{-0.0001686x}$	$f(x) = 0.7056e^{-0.00787x} + 0.2702e^{-0.000691x}$	$f(x) = 0.4022e^{-0.006308x} + 0.05879e^{-0.0005108x}$

Table 5: Bias % w.r.t. increase in shock's variability

Modified Populations	Cell size								
	Cell size	10	30	50	100	300	500	1000	5000	Overall Improvement (in % w.r.t baseline bias)
	0.5%	99	98	97	95	89	83	70	30 (18000)*	- 206 %
	1%	96	94	91	85	66	53	33	11 (5600)*	-
	2.5%	92	82	73	57	28	20	13 (1400)*	0.008	+ 64%
	6.75%	43	39	34	26	10	0.06	0.03	0.004	+ 87%

**Cell size for which estimated attenuation bias is below 90%*

Graph 4 – % Bias vs cell size across modified population



2.2 Modified population model - SSIV (shock int.)

Table 7 – Optimal cell size by range – Shock int. 0.5%

Second stage regressor: Cell size	First Stage regressor: 50	100	300	500	1000	3000	5000
50	76	130	97	98	-48	-307	3
100	160	54	-15	4	2	-28	-18
300	99	-89	14	-79	-12	-4	-2
500	99	-26	-23	-88	-4	-4	-2
1000	-27	53	-3	-5	2	-1	-3

Improvement w.r.t. baseline bias

- 1214%

+97%

MAX = 100%

MIN = - ∞

Table 7 – Optimal cell size by range – Shock int. 2.5%

Second stage regressor: Cell size	First Stage regressor: 50	100	300	500	1000	3000	5000
50	6	2	1	0	1	-3	-1
100	3	-2	1	-1	0	-2	1
300	1	0	0	-1	-1	0	0
500	4	1	1	-1	1	-1	-1
1000	0	1	1	-1	1	1	-1

$$f(x,y) = \alpha + \gamma_1 X + \gamma_2 Y + \gamma_3 X^2 + \gamma_4 XY$$

Coefficients (with 95% confidence bounds):

α = 23.43 (-0.3843, 47.24)
γ_1 = -32.58 (-59.35, -5.807)
γ_2 = -15.28 (-28.79, -1.76)
γ_3 = 11.1 (-9.214, 31.42)
γ_4 = 7.539 (-6.174, 21.25)

Goodness of fit: R-square: 0.6673

Panel data lenght = 2 (4 years)

Panel data lenght = 8 (16 years)

Baseline – Shock intensity 1%

Second stage regressor: Cell size	First Stage regressor: 50	100	300	500	1000	3000	5000
50	952	-162	-4	1	-4	4	6
100	-45	25	-7	-5	-8	2	3
300	-3	2	2	-2	-1	2	1
500	-3	6	2	-2	-1	1	3
1000	6	1	1	0	1	0	0

$$f(x,y) = \alpha + \gamma_1 X + \gamma_2 Y + \gamma_3 X^2 + \gamma_4 XY$$

Coefficients (with 95% confidence bounds):

α = 0.9291 (-0.997, 2.855)
γ_1 = -0.3301 (-2.047, 1.386)
γ_2 = -0.7622 (-2.413, 0.8883)
γ_3 = -0.08476 (-2.189, 2.02)
γ_4 = 0.3422 (-1.185, 1.87)

Goodness of fit: R-square: 0.6933

2.2 Modified population: FE (time)

Exponential fit for modified population model: average prop. of immigrants

Time = 2	Time = 4	Time = 8
Coefficients (with 95% confidence bounds): a = 0.6575 (0.5492, 0.7659) b = -0.003118 (-0.004032, -0.002204) c = 0.3298 (0.2146, 0.4449) d = -0.0003204 (-0.0004526, -0.0001882) Goodness of fit: SSE: 0.001732 R-square: 0.9987	Coefficients (with 95% confidence bounds): a = 0.7211 (0.625, 0.8172) b = -0.001822 (-0.002215, -0.001429) c = 0.2618 (0.1627, 0.361) d = -0.000172 (-0.0002532, -9.09e-05) Goodness of fit: SSE: 0.001548 R-square: 0.9989	Coefficients (with 95% confidence bounds): a = 0.6956 (0.6134, 0.7778) b = -0.01485 (-0.01852, -0.01117) c = 0.2365 (0.1435, 0.3294) d = -0.001537 (-0.002219, -0.0008551) Goodness of fit: SSE: 0.0006549 R-square: 0.999
$f(bias) = 0.6575e^{-0.003118x} + 0.3298e^{-0.0003204x}$	$f(bias) = 0.7211e^{-0.001822x} + 0.2618e^{-0.000172x}$	$f(bias) = 0.6956e^{-0.001485x} + 0.2365e^{-0.001537x}$

Graph – % Bias vs cell size across modified populations

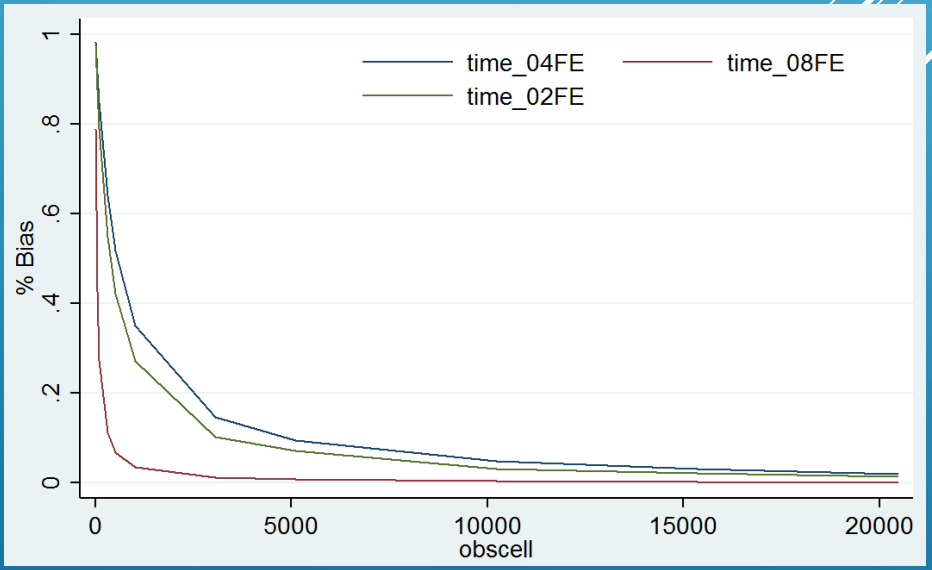


Table 5: Bias % w.r.t. increase in panel data length

Modified Populations: Cell size	10	50	100	300	500	1000	3000	5000	10000	15000	Overall Improvement (in % w.r.t 1 baseline bias)
T = 2	97	87	79	54	42	27	12 (3500)*	7	3	2	+ 28%
T = 4	98	91	84	63	51	34	16	11 (5600)*	4	3	-
T = 8	83	55	36	15	11 (550)*	0	0	0	0	0	+ 63%

**Cell size for which estimated attenuation bias is below 90%*

2.2 Modified population model - SSIV (Time)

Table – Optimal cell size by range – Time = 2

Second stage regressor \ First Stage regressor	100	300	500	1000	3000	5000
100	-19	-120	-11	-6	-12	-1
300	-7	-4	-5	-4	-1	1
500	-4	-3	-3	1	-2	-3
1000	-1	-3	-1	-1	-1	-1

Improvement w.r.t. baseline bias

- 564%

+99.99%

MAX = 100%

MIN = - ∞

Table – Optimal cell size by range – Time = 8

Second stage regressor \ First Stage regressor	100	300	500	1000	3000	5000
100	0	-1	0	1	0	0
300	-1	-1	0	0	0	0
500	0	-1	0	-1	0	0
1000	0	0	0	0	1	0

$$f(x,y) = \alpha + \gamma_1 X + \gamma_2 Y + \gamma_3 X^2 + \gamma_4 XY$$

Coefficients (with 95% confidence bounds):

α = 3.165 (-9.917, 16.25)
γ_1 = -1.669 (-14.55, 11.21)
γ_2 = -1.682 (-9.01, 5.646)
γ_3 = 0.2679 (-11.15, 11.68)
γ_4 = 0.9133 (-6.572, 8.399)

Goodness of fit:
SSE: 5357 R-square: 0.5974

Baseline – Time = 4

Second stage regressor \ First Stage regressor	100	300	500	1000	3000	5000
100	25	-7	-5	-8	2	3
300	2	2	-2	-1	2	1
500	6	2	-2	-1	1	3
1000	1	1	0	1	0	0

$$f(x,y) = \alpha + \gamma_1 X + \gamma_2 Y + \gamma_3 X^2 + \gamma_4 XY$$

Coefficients (with 95% confidence bounds):

α = 1.205e-14 (-0.233, 0.233)
γ_1 = -2.129e-14 (-0.2293, 0.2293)
γ_2 = -1.497e-14 (-0.1305, 0.1305)
γ_3 = 8.007e-15 (-0.2033, 0.2033)
γ_4 = 8.124e-15 (-0.1333, 0.1333)

Goodness of fit:
SSE: 1.699 R-square: 0.5708

3.1 – POLS comparison

Graph – POLS with FE

S.Rate	0.1	0.3	0.5	1	3	5
$\hat{\beta}^{FE}_{popul}$	.997	1	.999	1	.999	.999
s.e.	.0191	.011	.00854	.00604	.00349	.0027
$\hat{\beta}^{POLS}_{sample}$	.119	.285	.394	.568	.805	.872
s.e.	.0067	.00605	.00551	.00462	.00316	.00254
Obs Cell	69.7	209	348	696	2088	3480
s.e.	2.84	8.74	14.7	29.1	87.5	146

$y = 56{,}52e^{-0{,}009956x} - 63{,}05\ e^{-0{,}0003687x}$

Goodness of fit:
SSE: 688.1
R-square: 0.8503

Graph – POLS without FE

S.Rate	0.1	0.3	0.5	1	3	5
$\hat{\beta}^{FE}_{popul}$	.996	.996	.996	.997	.997	.997
s.e.	.00814	.0047	.00364	.00257	.00149	.00115
$\hat{\beta}^{POLS}_{sample}$	.524	.762	.848	.912	.967	.978
s.e.	.00636	.00428	.00345	.0025	.00147	.00114
Obs Cell	104	313	522	1043	3130	5216
s.e.	4.28	12.8	21.4	42.8	128	214

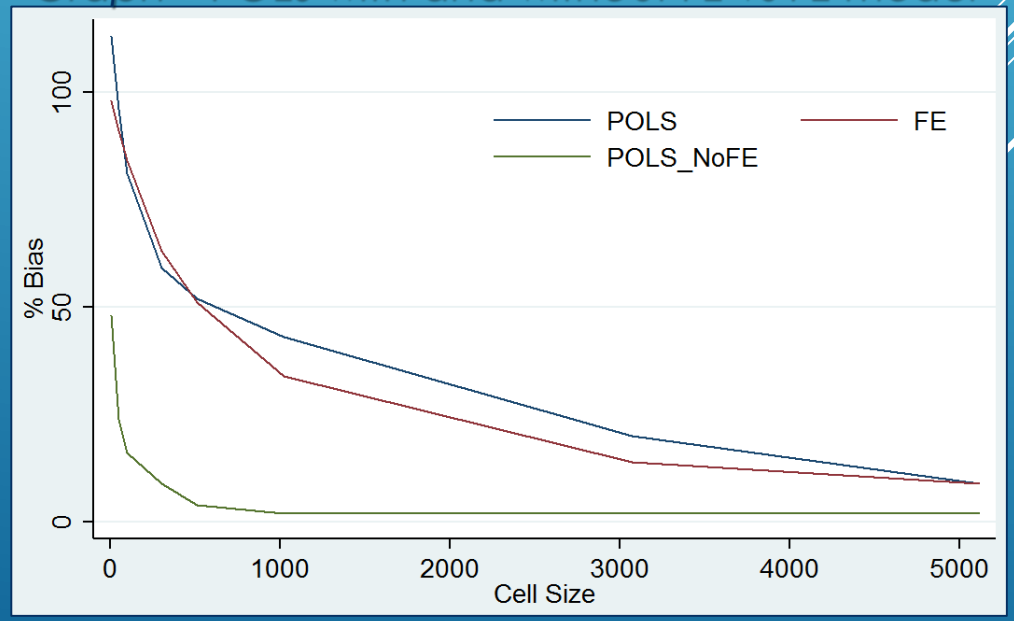
$y = 58{,}96e^{-0{,}004737x} - 12{,}3\ e^{-0{,}0003304x}$

Goodness of fit:
SSE: 3.475
R-square: 0.9979

Table 4 – Optimal cell size by range

Modified Populations	Cell size								
	Cell size	10	50	100	300	500	1024	3000	5000	Overall Improvement (in % w.r.t POLS With FE)
	POLS With FE	113	96	81	59	52	43	20	9	-
	POLS No FE	48	24	16	9	4	2	2	2	+350%
	F.E.	96	94	85	66	53	33	11	9	-

Graph – POLS with and without FE vs FE model



Conclusions

Table – Optimal Cell size for an approximate 90% bias reduction

Model	Baseline	Init. Migr. Prop.		Shock intensity		time	
	IMP = 0.1 SI = 1% TU = 4	*0.05*	*0.3*	*0.5*	*2.5%*	*2*	*8*
FE	5200	6650	3500	18000	1400	5000*	500
SSIV 2° Stg. variable/ intstrument	300/100 (200/50)	500/300	300/50	1000/300	50/50	300/300 (300/100)	50/50

**Estimated cell size for 90% bias reduction = 3500*

- SSIV is a more efficient variant
- Bigger yearly migratory shocks add more information to the regressor than a bigger initial share of migrants.
- Longer period of observation reduce the bias by extremely significant amount
- Presence of fixed effect causes a bigger attenuation bias

Thanks for your attention.