

Online Appendix

Childhood Trajectories of Neighbourhood Disadvantage and Educational Attainment: Towards a Longitudinal Machine-Learning Approach*

Maël Lecoursionnais, Haley McAvay

A Attrition analysis

To assess the effect of attrition on our results, we compared the distribution of key variables (high-school graduation, gender, foreign background) in two samples, one with all individuals born in 1989 and present in the registers in 1990 (sample A) and one with all individuals born in 1989 and present in the registers from 1990 to 2014. Sample A contains 116,744 individuals while sample B has 112,436. Table A1 provides the results of the analysis. The first two columns show the distribution of different variables for samples A and B. Looking at the first two columns, samples A and B show no substantial difference in the distribution of gender, high-school graduation and foreign background. In this case, attrition does not change the distribution of our variables, and ultimately does not affect statistical inference.

Tab. A1: Attrition analysis

	Sample	
	A	B
<i>Gender (in percentage)</i>		
Male	51.4	51.4
Female	48.6	48.6
<i>High-school graduation (in percentage)</i>		
Graduated	79.5	81.1
Did not graduate	20.5	18.9
<i>Foreign background (in percentage)</i>		
Native	81.2	82.8
Children of immigrants	18.8	17.2
<i>Sample size</i>	116,744	112,436

Source: Authors' calculations based on Swedish population registers, 1990-2017.

* This Online Appendix contains additional information regarding the article:
<https://www.comparativepopulationstudies.de/index.php/CPoS/article/view/752/478>

B Robustness checks

We use robustness checks to demonstrate that our results are not sensitive to the specific thresholds we imposed on our data set. These parameters include the radius which defines the maximum distance from which the k -nearest neighbours must live (default is 1000 metres), the estimation technique of the weights (default is BART), the life stage thresholds (default is 1, 6, and 11 years old) and the treatment quantile threshold (default is 0.8, meaning the top 20 percent unemployed neighbourhoods are considered treated). Table B1 displays different regression models with changes in those parameters. Six regression coefficients are indicated: the overall cumulative exposure spanning the entire observed period, the coefficients for life stages, and the interaction terms. Models in Table B1 show no major difference from the models presented in the results: changing the radius threshold to 2000 metres shows only little variation in the coefficients of interest. Without a radius threshold, we get smaller effects. This aligns with the idea that neighbourhood effects are more important on a local scale. Changing the quantile threshold to 0.7 or 0.9, instead of 0.8, does not significantly modify the coefficients.

Tab. B1: Robustness checks. Coefficients are log odds

Characteristics	Models						
	A	B	C	D	E	F	G
Cumulative exposure							
Overall	-0.027 (-0.003)	-0.021 (-0.002)			-0.028 (-0.004)	-0.028 (-0.003)	-0.031 (-0.004)
Period 1	-0.031 (-0.011)	-0.024 (-0.008)	-0.028 (-0.008)	-0.053 (-0.018)	-0.034 (-0.015)	-0.038 (-0.011)	-0.032 (-0.016)
Period 2	-0.009 (-0.014)	-0.008 (-0.010)	0.034 (-0.028)	-0.005 (-0.009)	-0.031 (-0.019)	0.003 (-0.014)	-0.002 (-0.020)
Period 3	-0.041 (-0.012)	-0.048 (-0.008)	-0.092 (-0.024)	-0.079 (-0.021)	-0.020 (-0.017)	-0.051 (-0.011)	-0.059 (-0.017)
Radius (default is 1000 metres)	2000	∞	def.	def.	def.	def.	def.
Estimation technique (default is BART)	def.	def.	def.	def.	PS	def.	def.
Life stage thresholds (default is 1/6/11)	def.	def.	1/10/13	1/4/13	def.	def.	def.
Quantile threshold (default is 0.8)	def.	def.	def.	def.	def.	0.7	0.9

Source: Authors' calculations based on Swedish population registers, 1990–2017.

C Sensitivity analysis

Our causal model assumes unconfoundedness, meaning that all variables affecting both the treatment and the outcome are observed and controlled for. While this assumption can never be entirely satisfied in observational studies, we use robustness checks and sensitivity analyses to assess the quality of our results. We use the E-value to evaluate the degree to which unobserved confounders would need to affect the outcome and treatment in order to null out the treatment effect (VanderWeele/Ding 2017; VanderWeele/Mathur 2020; VanderWeele et al. 2019).

The E-value is a synthesis measure that captures the “minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates” (VanderWeele/Ding 2017: 1). In other words, the larger the E-value, the more unobserved confounding would need to be in order to eliminate the observed association between the treatment and the outcome. The E-value does not require strong assumptions on unmeasured confounding, which makes its interpretation easy compared to other methods. E-values are reported on the risk ratio scale, which differs substantially from odds ratio when the prevalence of the outcome among the non-exposed group is high. This is the case in our study, as 88 percent of never-exposed students graduate from high school. We use the following equation to convert odds ratios to risk ratios (Zhang/Yu 1998):

$$RR = \frac{OR}{(1 - p_{ref}) + (p_{ref} * OR)}$$

with RR indicating the risk ratio, OR the odds ratio and p_{ref} the prevalence of the non-exposed group. As stated by Ding and VanderWeele (2016), for a given risk ratio RR, the E-value is

$$RR = \frac{OR}{(1 - p_{ref}) + (p_{ref} * OR)}$$

We report in Table C1 the E-values for 5 and 15 years of living in a disadvantaged neighbourhood, 5 years of exposure during early childhood, and 5 years of exposure during adolescence. The E-value corresponding to a 5-fold increase of β_1^w – an odds ratio of 0.977 – in the first model is 1.153 with a standard error of 0.004. It means that the association reported in Model 1 between the treatment and the outcome could be explained away by one or several unobserved confounders associated with the treatment and outcome at a risk ratio of 1.153 (1.149 for the lower confidence interval) each, after controlling for observed covariates. To put this into perspective, the risk ratio for the effect of having at least one parent with a Ph.D. degree on high-school graduation (compared to parents having finished in high school), i.e., the highest coefficient for a control variable in the first model, is 1.105 (4.884 in odds ratio). Even if the association between parents’ education and selection into

neighbourhood disadvantage is equally strong, it would still be insufficient to explain away the effect observed for five years of exposure to a disadvantaged neighbourhood, let alone for fifteen years of exposure. E-values of the estimate of five years of exposure during early childhood and adulthood are also higher than the strongest coefficient of Model 2.

Tab. C1: E-values for sensitivity analysis

	Coef.	2.5% CI	97.5% CI
5-year exposure			
Odds ratio	0.869	0.865	0.875
E-value	1.153	—	1.149
15-year exposure			
Odds ratio	0.658	0.654	0.662
E-value	1.319	—	1.316
5-year exposure during childhood			
Odds ratio	0.809	0.79	0.829
E-value	1.199	—	1.183
5-year exposure during adolescence			
Odds ratio	0.785	0.765	0.805
E-value	1.217	—	1.202

Source: Authors' calculations based on Swedish population registers, 1990-2017.

D Other tables

Among year-individuals belonging to the last quintile of the first principal component (i.e., the neighbourhood disadvantage measure), there are on average 25.0 percent of the 500 nearest neighbours who are unemployed, compared to 10.3 percent on average for other quintiles.

Tab. D1: Mean percentage value across all observed years of variables used in the PCA

Last quintile	Unemployment	Social assistance in %	Low income
Yes	25.0	25.9	32.6
No	10.3	5.04	16.3

Source: Authors' calculations based on Swedish population registers, 1990-2017.

Tab. D2: Regression output (model 1). Logistic regression with inverse probability weights. Coefficients are log odds

Model 1	Coef.	SE	p-value
Cumulative exposure to disadvantaged neighbourhood	−0.028	0.003	< 0.001
Gender			
Men	—	—	
Women	0.387	0.029	< 0.001
Number of moves	−0.18	0.009	< 0.001
Foreign background	—	—	
Native			
Children of immigrants	−0.161	0.039	< 0.001
Parents' highest educational attainment			
Compulsory schooling (less than 9 years)	−0.315	0.157	0.045
Compulsory schooling (9 years)	−0.502	0.065	< 0.001
Secondary education	—	—	
Post-secondary education (less than 2 years)	0.438	0.058	< 0.001
Post-secondary education (2+ years)	0.805	0.032	< 0.001
Graduate program	1.586	0.143	0.001
Observations		61,315	

Source: Authors' calculations based on Swedish population registers, 1990–2017.

Tab. D3: Regression output (model 2). Logistic regression with inverse probability weights

Model 2	Coef.	SE	p-value
Cumulative exposure to disadvantaged neighbourhood			
0-5 years old	-0.042	0.012	< 0.001
6-10 years old	-0.004	0.016	0.79
11-16 years old	-0.048	0.013	< 0.001
Gender			
Men	—	—	
Women	0.386	0.029	< 0.001
Number of moves	-0.188	0.009	< 0.001
Foreign background			
Native	—	—	
Children of immigrant	-0.158	0.029	< 0.001
Parents' highest educational attainment			
Compulsory schooling (less than 9 years)	-0.316	0.157	0.044
Compulsory schooling (9 years)	-0.500	0.065	< 0.001
Secondary education	—	—	
Post-secondary education (less than 2 years)	0.438	0.058	< 0.001
Post-secondary education (2+ years)	0.804	0.032	< 0.001
Graduate program	1.586	0.143	< 0.001
Observations		61,315	

Source: Authors' calculations based on Swedish population registers, 1990-2017.

E Inverse probability weighting

In this section, we expand on the application of inverse probability of treatment weighting (IPTW) to address the issue of time-varying confounding affected by past treatments. To explain its necessity, consider the directed acyclic graph (DAG) in Figure E1a, where each arrow represents an assumed causal effect. In this framework, time-invariant covariates (C) precede time-varying covariates (L) and neighbourhood history (A), while the outcome (Y) is measured subsequently.

Time-varying covariates are a function of past neighbourhood history and are assumed to affect both high school graduation and future neighbourhood selection. For example, level of parental economic resources predicts children's educational attainment and are an important determinant of future neighbourhood choice. At the same time, economic resources are affected by past neighbourhood trajectory.

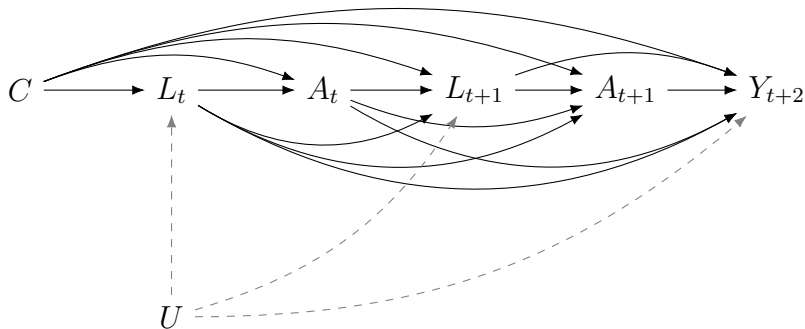
As depicted in Figure E1a, conditioning on a treatment-induced confounder like L_{t+1} in a conventional regression model has two key implications: (1) it blocks the causal pathway from earlier treatments to the outcome via the confounder (e.g., $A_t \rightarrow L_{t+1} \rightarrow \dots \rightarrow Y_{t+2}$), and (2) it introduces bias by opening a spurious path from earlier treatments to the outcome through unmeasured variables that affect both future time-varying confounders and the outcome (e.g., $A_t \rightarrow L_{t+1} \leftarrow U \rightarrow Y_{t+2}$).

Figure E1b shows a modified version of Figure E1a that illustrates how IPTW and MSM overcome problems of conventional regression methods. Weights are constructed such that the treatment is independent of past time-varying confounders, thus artificially removing the link between time-varying covariates and the treatment. In other words, the bias that arose from controlling for treatment-induced confounding is already removed in the weighted MSM, and additionally controlling for time-invariant confounders leads to a consistently unbiased estimation of the causal effect.

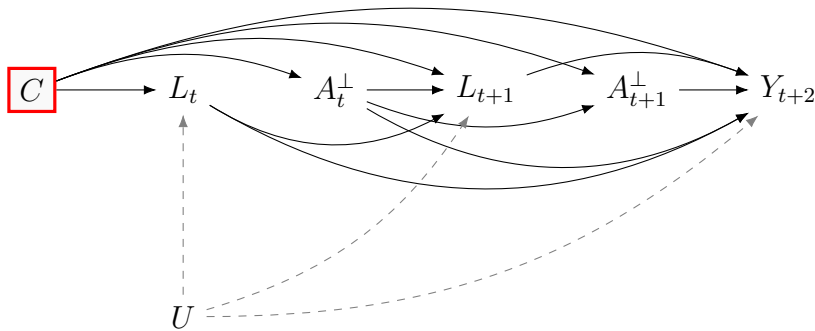
IPTW consists of weighting the sample by the individuals' inverse probability of treatment conditional on past treatments and confounders. The weights are constructed such that the reweighted probability of treatment equals 1. Lastly, weights are "stabilised" (Robins *et al.* 2000) to avoid high variance. We therefore have the following stabilised weight w_i :

$$\forall i, sw_i = \prod_{t=1}^T \frac{P(A_t = A_{i,t} | \bar{A}_{t-1} = \bar{a}_{i,t-1})}{P(A_t = A_{i,t} | \bar{A}_{t-1} = \bar{a}_{i,t-1}, \bar{X}_t = \bar{x}_{i,t})}.$$

Fig. E1: (a) DAG of neighbourhood effects from childhood to adulthood. (b) DAG of neighbourhood effects from childhood to adulthood with IPTW and MSM. The squared box represents conditioning. C represents time-invariant covariates, L time-varying ones, A neighbourhood disadvantage, Y high school graduation, and U unobserved variables



(a)



(b)

Source: Authors' calculations based on Swedish population registers, 1990-2017.

References

- Ding, Peng; VanderWeele Tyler J. 2016: Sensitivity Analysis Without Assumptions. In: *Epidemiology* 27,3: 368-377. <https://doi.org/10.1097/EDE.0000000000000457>
- Robins, James M.; Hernán, Miguel Ángel; Brumback, Babette 2000: Marginal Structural Models and Causal Inference in Epidemiology. In: *Epidemiology* 11,5: 550-560. <https://doi.org/10.1097/00001648-200009000-00011>
- VanderWeele, Tyler J.; Ding, Peng 2017: Sensitivity Analysis in Observational Research: Introducing the E-Value. In: *Annals of Internal Medicine* 167,4: 268-284. <https://doi.org/10.7326/M16-2607>

- VanderWeele, Tyler J.; Mathur, Maya B.; Ding, Peng* 2019: Correcting Misinterpretations of the E-Value. In: *Annals of Internal Medicine* 170,2: 131-132. <https://doi.org/10.7326/M18-3112>
- VanderWeele, Tyler J.; Mathur, Maya B.* 2020: Commentary: Developing Best-Practice Guidelines for the Reporting of E-Values. In: *International Journal of Epidemiology* 49,5: 1495-1497. <https://doi.org/10.1093/ije/dyaa094>
- Zhang, Jun; Yu, Kai F.* 1998: What's the Relative Risk? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. In: *JAMA* 280,19: 1690-1691. <https://doi.org/10.1001/jama.280.19.1690>

Dr. Maël Lecoursonnais (✉). Linköping University. Linköping, Sweden.
E-mail: mael.lecoursonnais@liu.se
URL: <https://liu.se/en/employee/maele63>

Prof. Dr. Haley McAvay. The London School of Economics and Political Science. London, United Kingdom.
E-mail: h.e.mcavay@lse.ac.uk
URL: <https://www.lse.ac.uk/people/haley-mcavay>

Comparative Population Studies

www.comparativepopulationstudies.de

ISSN: 1869-8980 (Print) – 1869-8999 (Internet)

Published by

Federal Institute for Population Research
(BiB)
65180 Wiesbaden / Germany



2026

Editor

Prof. Dr. Roland Rau
Prof. Dr. Heike Trappe

Managing Editor

Dr. Katrin Schiefer

Editorial Assistant

Beatriz Feiler-Fuchs
Wiebke Hamann

Layout

Beatriz Feiler-Fuchs

E-mail: cpos@bib.bund.de

Scientific Advisory Board

Kieron Barclay (Stockholm)
Ridhi Kashyap (Oxford)
Anne-Kristin Kuhnt (Rostock)
Mathias Lerch (Lausanne)
Eleonora Mussino (Umeå)
Natalie Nitsche (Canberra)
Alyson van Raalte (Rostock)
Pia S. Schober (Tübingen)
Sergi Vidal (Barcelona)