Do subway openings reduce air pollution? A replication exercise

Michael Wiebe

April 11, 2024

Abstract

Gendron-Carrier et al. (2022) studies the effect of subway openings on urban air pollution. The authors find a null average effect, but a negative effect in cities with high initial pollution. In this comment, I perform several robustness checks on the negative effect for high-pollution cities, and repeat the main analyses for low-pollution cities. I show that the main finding for high-pollution cities is robust, and find mixed results for low-pollution cities. I implement an alternative back-of-the-envelope calculation for the effect of subway openings on infant mortality, and find a smaller number of averted deaths.

1 Introduction

Gendron-Carrier et al. (2022), henceforth GGPT, studies the effect of subway openings on urban air pollution. They use a global sample of new subway openings over 2001-2016 combined with satellite data on Aerosol Optical Depth (AOD) as the measure of pollution. GGPT uses event study and difference-in-differences specifications, with treatment variation coming from the staggered timing of subway openings. The sample includes 58 cities that opened new subways, as well as an extended sample of control cities.

If people switch from pollution-emitting cars to taking the subway, then subway openings should decrease air pollution. GGPT find a null average effect, but a negative effect in cities with high initial pollution. They show that this subgroup effect is robust across different specifications, grows larger over time, and decays with distance from the city center. They also provide suggestive evidence that substituting car traffic is the mechanism in reducing pollution, since the effect is larger in cities with more subway ridership.

GGPT describe their main result as follows: "For the 10,896 city months in high AOD cities, AOD averages 0.66. From Table 4, the subway effect is about -0.028. Thus, our benchmark estimate for the effect of subways on AOD in polluted cities is about a 4 percent decline." (p.184) GGPT appear to be

averaging the estimates in Columns 5 and 6 of Table 4, which are -0.0270 and -0.0284, with corresponding p-values 0.077 and 0.013.[1]

In this comment, I perform several robustness checks on results in GGPT. I obtained the original code and data from the replication archive (Gendron-Carrier et al., 2021). For the negative effect in high-pollution cities, I test for robustness to alternative definitions of "high-pollution", using the 40th and 60th percentiles of initial pollution instead of the median; and an alternative measure of initial pollution, using AOD from year 2001 instead of 2000. I also use a continouous interaction effect, instead of the binary interaction in Table 3, Column 8. Finally, I rerun the event study separately by high- and low-pollution cities, since GGPT only report the event study for the full sample. I find that the main result for high-pollution cities is robust.

I repeat the main analyses for low-pollution cities, since the authors report a positive estimate (indicating that subways increase pollution) with a t-statistic of 1.5, but do not investigate further. I find mixed results. Moreover, I investigate the ridership mechanism, and find that high-pollution cities also have more subway riders, indicating that the heterogeneity by initial pollution and ridership is consistent. Finally, I revisit the back-of-the-envelope calculation for the effect of subway openings on infant mortality. I find that the authors underestimated the number of averted deaths: 25 averted deaths per year compared to 22.5 as reported in the original paper. Taking an alternative approach using the city-specific treatment effects instead of the average effect, I find a smaller number of averted deaths (15 per year).

2 Computational reproducibility

The original code is written for a Windows server, and uses backslashes in the filepaths. Since I am using Linux, I have to change the paths to instead use forward-slashes. I also have to translate the shell commands to the Linux equivalent. With these changes, I am able to exactly reproduce the main results using the original data and code.

3 Robustness replication

3.1 Is the effect in high-pollution cities robust?

GGPT's main finding is that subways reduce pollution in cities with above-median initial pollution. This finding is plausible, since we would expect it to be easier to reduce pollution in cities where pollution is high. But given the lack of pre-registration, we should approach this evidence with a skeptical prior. GGPT present 8 interaction effects in Table 3, and actually run 15 interactions in their

[1]Since the AEA style guide omits stars indicating statistical significance, I rerun the Table 4 results to obtain these p-values.

code. Hence, they had many opportunities to find a dimension that produced statistical significance, leading to a multiple testing problem.[2]

To check whether this finding is reliable, I perform several robustness checks. First, I vary the threshold used in defining high-pollution cities. GGPT naturally use the median to define a dummy variable for high-pollution cities. I rerun their analysis using the 40th and 60th percentiles, and find estimates with similar magnitude (see Table 1 and Table 2). However, using a 5% significance level, none of the estimates in Table 1 are statistically significant, and only Column 6 in Table 2 is statistically significant.

Table 1: Replication of Table 4: Defining High AOD Cities with P40

	(1)	(2)	(3)	(4)	(5)	(6)
post	-0.0237	-0.0185	-0.0196	-0.0221	-0.0215	-0.0142
	(0.0179)	(0.0190)	(0.0132)	(0.0132)	(0.0132)	(0.0107)
satellite	N	Y	Y	Y	Y	Y
cont.×year	N	Y	Y	Y	Y	Y
city×cal. mo.	N	N	Y	Y	Y	Y
climate × cont.	N	N	N	N	Y	Y
Mean AOD	0.61	0.61	0.61	0.61	0.61	0.42
bootstrap p-value	0.178	0.334	0.147	0.096	0.111	0.200
R^2	0.43	0.45	0.70	0.71	0.71	0.75
# events	34	34	34	34	34	40
# cities	34	34	34	34	34	501
N	12730	12730	12730	12730	12730	187708

Note: High AOD cities are defined as having initial AOD levels above the 40th percentile. Dependent variable is mean AOD in a 10km disk with centroid in the city center. Standard errors clustered at the city level are in parentheses.

Second, in Table 3, I vary the year used in defining initial high-pollution cities. The first subway opening is in 2002 (see Table A7), while the pollution data starts in 2000. GGPT define initial pollution using data from year 2000. When I rerun their analysis using initial pollution defined in 2001, I find similar results. As with the original Table 4, the point estimate is significant at the 5% level only in Column 6.

Third, instead of estimating the interaction effect with a dummy variable for above-median initial pollution, in Table 4 I use a continuous measure of initial pollution. I find an effect of similar magnitude, though not statistically significant at the 5% level (see Column 1).[3] When restricting the sample to high-pollution cities and repeating the continuous interaction in Column 2, I find no relationship. So the data does not support a simple story of "subway openings reduce pollution more in more-polluted cities" (or at least, there is not enough variation in this subsample to detect it). For reference, I also present

[2]See Gelman and Loken (2014) for a discussion of the problem of the 'garden of forking paths'.

[3]Using this continuous estimate, moving from the 10th to 90th percentile of initial pollution reduces pollution by $-0.0733 * (0.73 - 0.19) = -0.04$. The effect using the author's binary treatment varible is -0.028.

Table 2: Replication of Table 4: Defining High AOD Cities with P60

	(1)	(2)	(3)	(4)	(5)	(6)
post	-0.0296	-0.0225	-0.0213	-0.0246	-0.0244	-0.0300**
	(0.0225)	(0.0232)	(0.0172)	(0.0173)	(0.0174)	(0.0134)
satellite	N	Y	Y	Y	Y	Y
cont.×year	N	Y	Y	Y	Y	Y
city×cal. mo.	N	N	Y	Y	Y	Y
climate × cont.	N	N	N	N	Y	Y
Mean AOD	0.73	0.73	0.73	0.73	0.73	0.43
bootstrap p-value	0.173	0.326	0.216	0.167	0.171	0.040
R^2	0.17	0.19	0.55	0.56	0.56	0.75
# events	23	23	23	23	23	28
# cities	23	23	23	23	23	489
N	8574	8574	8574	8574	8574	183157

Note: High AOD cities are defined as having initial AOD levels above the 60th percentile. Dependent variable is mean AOD in a 10km disk with centroid in the city center. Standard errors clustered at the city level are in parentheses.

Table 3: Replication of Table 4: initial AOD from 2001

	(1)	(2)	(3)	(4)	(5)	(6)
post	-0.0279	-0.0255	-0.0243	-0.0268*	-0.0265*	-0.0262**
	(0.0209)	(0.0209)	(0.0145)	(0.0144)	(0.0145)	(0.0133)
satellite	N	Y	Y	Y	Y	Y
cont.×year	N	Y	Y	Y	Y	Y
city×cal. mo.	N	N	Y	Y	Y	Y
climate × cont.	N	N	N	N	Y	Y
Mean AOD	0.66	0.66	0.66	0.66	0.66	0.43
bootstrap p-value	0.162	0.205	0.097	0.069	0.073	0.052
R^2	0.33	0.35	0.64	0.65	0.65	0.75
# events	29	29	29	29	29	29
# cities	29	29	29	29	29	490
N	10874	10874	10874	10874	10874	183526

Note: This table uses initial AOD from year 2001 when defining high-pollution cities. The original table used year 2000 AOD. Dependent variable is mean AOD in a 10km disk with centroid in the city center. Standard errors clustered at the city level are in parentheses.

this continuous interaction for the low-pollution subgroup in Column 3. Here the relationship is positive but not statistically significant.

Table 4: Replication of Table 3, Column 8: Continuous measure of AOD

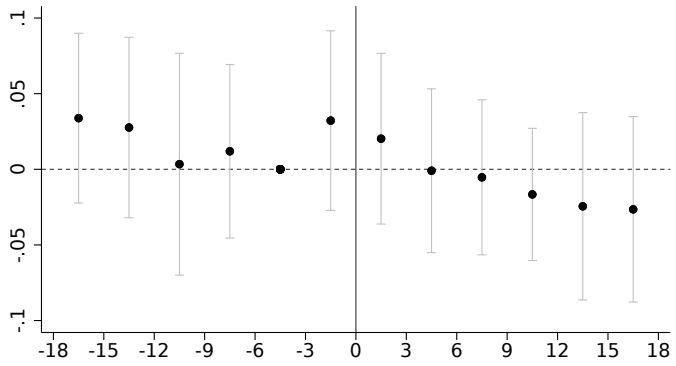
	(1)	(2)	(3)
post	0.0232	-0.0299	-0.0190
	(0.0167)	(0.0613)	(0.0244)
post $\times$ x	-0.0733	0.0047	0.0931
	(0.0451)	(0.1056)	(0.1070)
satellite	Y	Y	Y
cont.$\times$year	Y	Y	Y
city$\times$cal. mo.	Y	Y	Y
climate $\times$ cont.	Y	Y	Y
Mean AOD	0.46	0.66	0.25
R^2	0.80	0.66	0.68
# events	58	29	29
# cities	58	29	29
N	21806	10896	10910

Note: This regression interacts the Post dummy with a continuous measure of initial AOD. Column 1 uses the full sample, Column 2 restricts to high pollution cities, and Column 3 restricts to low pollution cities. Dependent variable is mean AOD in a 10km disk with centroid in the city center. Standard errors clustered at the city level are in parentheses.

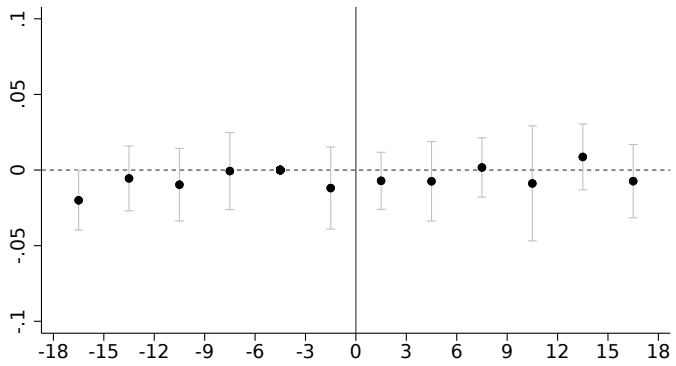
Fourth, I rerun the event study graph for high- and low-pollution cities. GGPT show this graph for the full sample to provide evidence against pre-trends, but do not run event studies separately by subgroup. Figure 1 shows a slight downward post-treatment trend in the high-pollution sample, corresponding to the negative average effect in that subsample. There is no post-treatment trend in the low-pollution sample. For both subsamples, the confidence intervals overlap zero.

Overall, the original findings for high-pollution cities are stable, but the precision of the estimates is not consistent, with p-values above and below 0.1.

Figure 1: Replication of Figure 4: event study by subsample



(a) High pollution



(b) Low pollution

Note: Event study coefficients for the high- and low-pollution samples, corresponding to the original Figure 4. The samples are split by median AOD in 2000. Coefficients are estimated using three-month bins; the reference bin is $\{-6, -5, -4\}$. Standard errors are clustered at the city level.

3.2 Do subways increase pollution in low-pollution cities?

GGPT find that subway openings actually increase pollution in low-pollution cities. The original Table 3, Column 8, reports a coefficient on the *Post* variable of 0.0126 with a standard error of 0.0083 (t-statistic=1.5). This effect is almost half the size of the pollution reduction in high-pollution cities, and would seem to warrant investigation. However, GGPT dismiss this effect on the basis of statistical insignificance.[4] This is unsatisfactory, because the negative effect in high-pollution cities is also nonsignificant if we correct for multiple testing with the Benjamini-Hochberg procedure.[5] To investigate whether the positive effect is robust, I replicate Tables 4-6 using low-pollution cities (instead of high-pollution cities as in the paper).

Table 5 replicates the paper's original Table 4 for low pollution cities. For high pollution cities, GGPT found that the point estimate was fairly robust to the addition of control variables. Here for low pollution cities, we find a point estimate that is fairly stable for the first five columns, but the magnitude is three times smaller than the coefficient from the original Table 3, Column 8. This change in magnitude appears to be driven by the smaller sample used in the subsample regression, whereas the original Table 3 uses the full sample and an interaction effect. Column 6, including the non-subway cities, has a larger effect size, perhaps indicating that the sample composition influences how the control variables affect estimation of the treatment effect. The estimates are not statistically significant at the 5% level.

Table 6 shows dynamic treatment effects for low pollution cities. For high pollution cities, GGPT found that treatment effects grow over time and persist, consistent with the mechanism of subway ridership displacing car traffic. Here for low pollution cities, we find inconsistent results, with some time periods having large positive effects, and others having equally large negative effects. This could be explained by changes in the sample: as the time period is extended, the number of treated cities decreases. With heterogeneity in treatment effects (see original Figure 5), the average effect on a different sample will be different.

Table 7 shows spatial decay. For high pollution cities, GGPT found that treatment effects shrink as distance from the city center increases. Here for low pollution cities, we find the opposite effect: subways increase pollution slightly more as distance from the city center grows. Although the difference in magnitude is small, one possible explanation is that new subways induce immigration to the suburbs in low-pollution cities.

Overall, the results from the low-pollution cities are mixed. The effect size

[4] In the abstract of the paper, GGPT say: "For less polluted cities, the effect is indistinguishable from zero."

[5] Table 3 reports eight separate interaction effects, with seven more included in the code. Focusing only on the reported interaction effects, I calculate p-values (in increasing order) of: 0.00654, 0.15, 0.3, 0.4, 0.9, 0.9, 0.9. The corresponding Benjamini-Hochberg critical values with $\alpha = 0.05$ and $m = 8$ tests are: 0.00625, 0.0125, 0.01875, 0.025, 0.03125, 0.0375, 0.04375, 0.05. Since each p-value is larger than the corresponding critical value, we fail to reject the null hypothesis in every case.

Table 5: Replication of Table 4: effect of subway openings in low-pollution cities

	(1)	(2)	(3)	(4)	(5)	(6)
post	0.0035	0.0032	0.0039	0.0040	0.0045	0.0101
	(0.0089)	(0.0090)	(0.0069)	(0.0072)	(0.0074)	(0.0067)
satellite	N	Y	Y	Y	Y	Y
cont.×year	N	Y	Y	Y	Y	Y
city×cal. mo.	N	N	Y	Y	Y	Y
climate × cont.	N	N	N	N	Y	Y
Mean AOD	0.25	0.25	0.25	0.25	0.25	0.40
bootstrap p-value	0.707	0.739	0.602	0.597	0.553	0.135
R^2	0.32	0.33	0.67	0.67	0.68	0.75
# events	29	29	29	29	29	29
# cities	29	29	29	29	29	490
N	10910	10910	10910	10910	10910	183562

Note: This table restricts the sample to cities with below-median initial AOD. Dependent variable is mean AOD in a 10km disk with centroid in the city center. Standard errors clustered at the city level are in parentheses.

Table 6: Replication of Table 5: long-run effects in low-pollution cities

	(1)	(2)	(3)	(4)	(5)	(6)
Panel a.						
1-12 months post	0.0047	0.0086	0.0043	0.0044	0.0010	0.0042
	(0.0083)	(0.0070)	(0.0080)	(0.0071)	(0.0093)	(0.0080)
13-24 months post	0.0057	0.0145**	0.0027	0.0062	-0.0057	0.0050
	(0.0075)	(0.0073)	(0.0072)	(0.0056)	(0.0085)	(0.0060)
25-36 months post			-0.0092	-0.0059	-0.0147*	-0.0046
			(0.0058)	(0.0056)	(0.0077)	(0.0063)
37-48 months post					-0.0044	0.0074
					(0.0116)	(0.0092)
satellite	Y	Y	Y	Y	Y	Y
cont.×year	Y	Y	Y	Y	Y	Y
city×cal. mo.	Y	Y	Y	Y	Y	Y
climate × cont.	Y	Y	Y	Y	Y	Y
Mean AOD	0.25	0.40	0.24	0.40	0.23	0.40
R^2	0.68	0.75	0.66	0.75	0.67	0.75
# events	28	28	25	25	21	21
# cities	28	489	25	486	21	482
N	10515	183167	9378	182030	7831	180483
Panel b.						
average post	0.0051	0.0115*	-0.0000	0.0016	-0.0043	0.0030
	(0.0074)	(0.0065)	(0.0065)	(0.0051)	(0.0082)	(0.0060)
bootstrap p-value	0.510	0.093	0.994	0.763	0.618	0.625

Note: This table restricts the sample to cities with below-median initial AOD. Dependent variable is mean AOD in a 10km disk with centroid in the city center. Standard errors clustered at the city level are in parentheses.

Table 7: Replication of Table 6: spatial decay of subway effects in low-pollution cities

	(1)	(2)	(3)	(4)	(5)	(6)
post	0.0045	0.0101	0.0069	0.0124*	0.0071	0.0121*
	(0.0074)	(0.0067)	(0.0082)	(0.0072)	(0.0078)	(0.0069)
satellite	Y	Y	Y	Y	Y	Y
cont.×year	Y	Y	Y	Y	Y	Y
city×cal. mo.	Y	Y	Y	Y	Y	Y
climate × cont.	Y	Y	Y	Y	Y	Y
Mean AOD	0.25	0.40	0.23	0.38	0.22	0.36
bootstrap p-value	0.553	0.135	0.425	0.101	0.356	0.080
R^2	0.68	0.75	0.68	0.81	0.71	0.82
# events	29	29	29	29	29	29
# cities	29	490	29	490	29	490
N	10910	183562	10910	183460	10910	183444

Note: This table restricts the sample to cities with below-median initial AOD. Odd columns restrict the sample to the 29 low pollution cities, while even columns include all non-subway cities. The dependent variable in columns 1 and 2 is mean AOD in a 10km disk with centroid in the city center. In columns 3 and 4 it is AOD in a donut 10-25km from the city center. In columns 5 and 6 it is AOD in a donut 25-50km from the city center. Standard errors clustered at the city level are in parentheses.

is smaller when using a subsample regression, the dynamic treatment effects are noisy, and the spatial effects have an unintuitive pattern (which is perhaps to be expected for a positive effect of subways on pollution).

3.3 The ridership mechanism

GGPT show in the original Table 8 that pollution decreases more in cities with above-median ridership (measured 12 months after the subway opening). Of the original 58 cities, 42 have ridership data, and GGPT measure ridership in both total and per-capita terms. Since the main finding is that pollution decreases more in high pollution cities, this raises the question of whether the high-pollution cities are the same as the high-ridership cities.

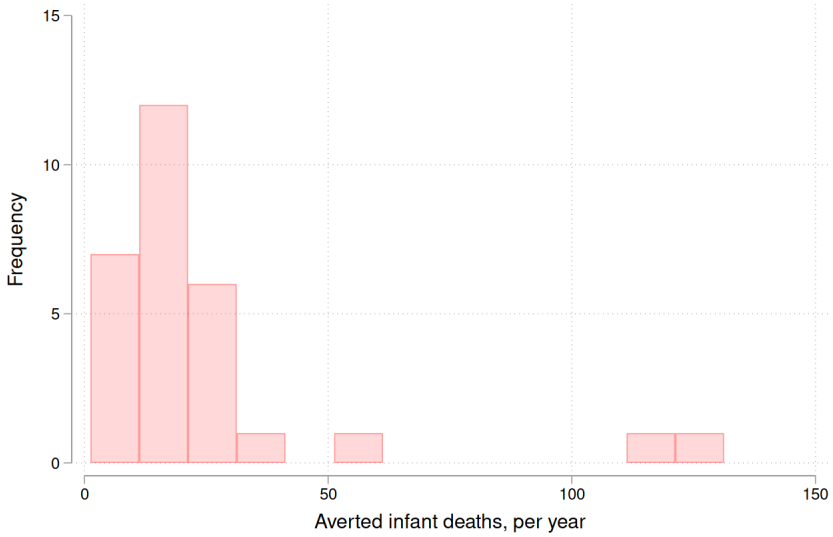
Using the per-capita ridership variable (corresponding to the original Table 8, Column 3), I find that 35/42 cities are both high-ridership and high-pollution. Four cities are high pollution but low ridership, and three are high ridership but low pollution. Using the total ridership variable (corresponding to Table 8, Column 4), I find that 25/42 cities are both high-ridership and high-pollution. Nine cities are high pollution but low ridership, and eight are high ridership but low pollution.

Since the high-ridership cities mostly overlap with the high-pollution cities, the finding of decreased pollution in both subgroups is consistent.

3.4 Averted deaths

Finally, I revisit GGPT's back-of-the-envelope calculations that translate the effect of subways on pollution into an effect on infant mortality. GGPT use results from the literature for the effect of PM10 on infant mortality, and convert the effect of subway openings on AOD into an effect on PM10. Using an average population of 5.3 million and a global average birthrate of 2%, the paper reports that an average subway opening in a high pollution city averts 34 infant deaths per year.[6] When using city-level population and country-level birthrate data, they report 22.5 infant deaths averted per year.[7]

Figure 2: Distribution of averted deaths, by city (average treatment effect)



Note: Averted deaths are calculated using the average treatment effect, city-level population, and country-level birthrate. The sample is restricted to high-pollution cities.

To understand the variation underlying this average effect, I plot the distribution of averted deaths in Figure 2. We can see that two cities, Delhi and Mumbai, have over 100 deaths averted. These are the largest cities in the sample, and also have the first and third largest birthrates, respectively. Since this calculation uses the average treatment effect of -0.028 for high pollution cities,

[6]Note that they use the global average birth rate of 2% from the World Bank, but the birthrate in the high pollution sample is 0.013. Using the sample birthrate, we get 22 deaths averted: 3.2 * (10/100000) * 5300000 * 0.013 = 22, using the formula: 3.2 micrograms/m^3 of PM10 averted * 10 infant deaths per 100,000 births per microgram/m^3 of PM10 * number of births = number of infant deaths averted.

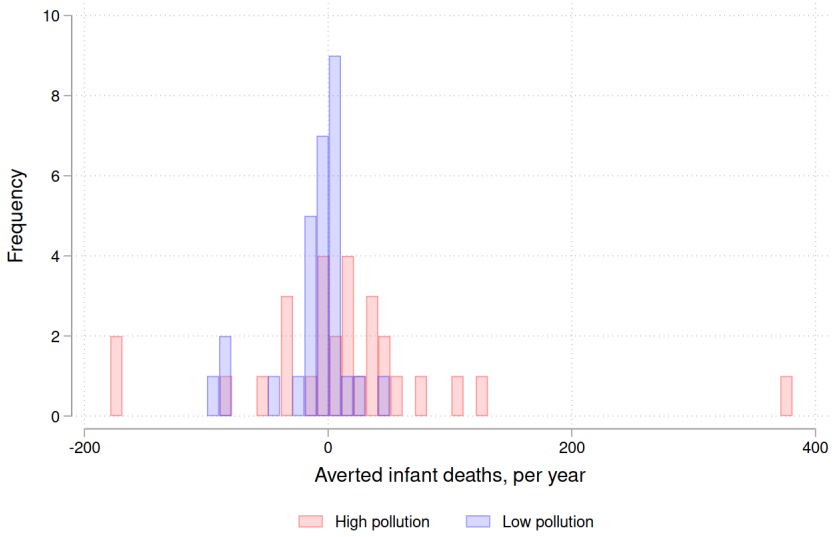
[7]The paper reports an estimate of 22.5, but this is due to an error in the code, which incorrectly multiplies by 2.9 instead of 3.2 = 0.028*114.6, where 0.028 is the effect size in AOD and 114.6 is the exchange rate from AOD to PM10. The correct number is 25 deaths averted.

the variance in averted deaths is entirely driven by variation in population and birthrates.

GGPT note that using city-level population data makes the back-of-the-envelope calculation more precise (compared to using average population). Similarly, we can use the city-level treatment effects to further improve precision. As shown in the paper's Figure 5, the majority of the city-specific treatment effects are significant at the 5% level, and the treatment effects that are non-significant have small point estimates. Hence, the city-specific estimates should be informative for this exercise.[8]

I calculate averted infant deaths for each city, using city-specific treatment effects along with the city-level population and country-level birthrate data. The distribution of averted deaths is shown in Figure 3. For comparison, I also include the estimates for the low-pollution cities. As we can see, there is quite a range of estimates, especially within the high-pollution cities. In Delhi, 371 deaths are averted per year, while in Chongqing and Mumbai, 179 and 170 deaths are *caused* per year, respectively.

Figure 3: Distribution of averted deaths, by city (city-specific treatment effect)



Note: Averted deaths are calculated using city-specific treatment effects, city-level population, and country-level birthrate. The sample includes both high- and low-pollution cities.

Averaging by subgroup, we find 15 infant deaths averted per year in high-pollution cities, and 11 deaths caused per year in low-pollution cities. The estimates of averted deaths are quite sensitive to the correlation between the city-level treatment effect, population, and birth rate. For example, Delhi has a medium treatment effect (-0.08, se=0.009) combined with a high population

[8] Arguably, the city-specific treatment effects should be estimated with a shrinkage estimator to reduce the influence of sampling variation.

and birthrate, and hence contributes substantially to the average in the high pollution subgroup. Similarly, Chennai has a medium treatment effect (0.05, se=0.01) and large population, and contributes the most to the average in the low pollution subgroup.

GGPT calculate the dollar value of averted deaths at $1 billion per year. But this estimate is based on 22.5 deaths averted per year. If we use 15 deaths averted per year (from the city-specific treatment effects), then the dollar value should be correspondingly smaller, by about 30%. Moreover, if we take seriously the city-specific estimates from the low pollution cities, then we also have to account for the costs of deaths caused (11 per year).

Using the city-specific treatment effects seems like the correct approach for estimating averted deaths in this particular sample of cities. However, when thinking about external validity and applying these results to a new (high pollution) city, averted deaths should be calculated using the average treatment effect (-0.028) and that city's population and birthrate data.

4 Conclusion

Overall, my replication exercise supports the finding of the paper that subway openings reduce air pollution. While the authors test for many interaction effects, the negative effect for the subgroup of high-pollution cities is in fact robust. The authors downplay the positive effect for the subgroup of low-pollution cities. I repeat the main analysis for low-pollution cities and find a mixed pattern of results. Finally, I show that by using the city-specific treatment effects to calculate averted deaths instead of the average effect, I find a smaller monetary value of pollution reduction.

References

Gelman, A. and E. Loken (2014). The statistical crisis in science. *American Scientist 102*.

Gendron-Carrier, N., M. Gonzalez-Navarro, S. Polloni, and M. A. Turner (2021). Data and Code for: Subways and Urban Air Pollution. American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor].

Gendron-Carrier, N., M. Gonzalez-Navarro, S. Polloni, and M. A. Turner (2022, January). Subways and urban air pollution. *American Economic Journal: Applied Economics 14*(1), 164–96.