Sewers and Urbanization in the Developing World*

Sean McCulloch,† Matthew Schaelling,‡
Matthew A. Turner,§ Toru Kitagawa¶

16 March 2025

Abstract: *We investigate the effects of sewer access on neighborhood characteristics in developing world cities. Because it is more difficult to move sewage uphill than downhill, otherwise similar neighborhoods on opposite sides of drainage basin divides may face different costs of sewer access. We exploit this intuition to identify the effect of sewer access by comparing outcomes for neighborhoods on opposite sides of drainage basin divides. We estimate the effect of sewer access on census tract population density, literacy, and income for Brazil, Colombia, South Africa, Jordan, and Tanzania. On average, sewer access has a large effect on population density and almost none on demographics. These estimates imply that sewer networks are often as important for the economic geography of cities as transportation networks.*

JEL: O18, R3, L97, N11
Keywords: Sewers, Urbanization, Infrastructure

*We are grateful to Victoria Delbridge and IGC country staff in South Africa, Pakistan, Jordan, Ghana and Zambia for help with data collection. We are also grateful for helpful comments from Alex Rothenberg and from seminar participants at the University of Wisconsin, the European and North American meetings of the Urban Economics Association, and the meetings of the Latin American and China Urban Economics Associations. This research was supported by IGC grants BRA-23005 and XXX-23154 and by the Brown University PSTC, which receives funding from the NIH, for training support (T32 HD007338) and for general support (P2C HD041020). Any errors are our responsibility alone.

†Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: sean_mcculloch@brown.edu.

‡Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: matthew_schaelling@brown.edu.

§Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: matthew_turner@brown.edu. Also affiliated with PERC, IGC, NBER, PSTC, S4.

¶Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: toru_kitagawa@brown.edu.

1 Introduction

We investigate the effects of sewer access on census tract population density, literacy rate, and mean income in developing world cities. Our estimates are based on a quasi-experimental research design that derives from principles of wastewater engineering. Because it is more difficult to move sewage uphill than downhill, otherwise similar census tracts on opposite sides of drainage basin divides sometimes face different costs of sewer access. We use this intuition, and census tract level data, to estimate the effects of treating census tracts with better sewer access. We identify treatment effects by comparing rates of sewer access and outcomes for census tracts on opposite sides of drainage basin divides in Brazil, Colombia, South Africa, Jordan, and Tanzania. We use these estimates to evaluate the impact of sewers on a sample of large cities in these same countries.

We provide two types of estimates. The first is a conventional TSLS/IV. At the census tract level, our treatment variable, share of households with sewer access, is continuous. This means that, outside of a homogeneous treatment effect framework, the TSLS/IV LATE involves an average of treatment effects that is not obviously of economic interest. To address this issue, we note that at the parcel level our treatment effect is binary. A parcel either has sewer access or not. We exploit this observation to estimate a parcel level MTE/LIV model with census tract level data by using a small variance approximation (Chesher, 1991). Unlike the TSLS/IV estimand, the interpretation of the MTE/LIV based estimand is simple and economically meaningful. It is the census tract average of the parcel level effect of sewer access. In practice, the magnitude of the TSLS/IV and MTE/LIV estimates are about the same.

Our preferred estimates indicate that increasing the number of sewer connections in a census tract by 1% increases population density by about 6%. Using this estimate to evaluate small counterfactual sewer expansions suggests that expanding sewer networks has about equal, but opposite, effects on urban density as large expansions of transportation networks. We also find that sewer access has only small effects on tract mean income and literacy. This suggests that sewers make a neighborhood more attractive to people like those who already live there. Sewer access does not lead to the displacement of poor slum dwellers by more affluent newcomers or to other dramatic shifts in tract demographics.

The economic logic of cities is simple. We are more productive if we work at higher densities than we can tolerate in our residences. Cities arise as the joint arrangement of work and residence locations that allows higher employment and lower residential density. Stated in this way it is natural that researchers focus their attention on how transportation infrastructure affects the development of cities. However, our willingness to tolerate density is also fundamental to how cities are organized, and the ability of infrastructure to facilitate density, as opposed to mobility, has received little attention from researchers. Sewer access has obvious implications for our willingness to tolerate population density, and so this paper begins the study of the importance of infrastructure that facilitates population density for the development of cities.

According to the World Bank, about one third of the world's urban population did not have access to safely managed sanitation facilities in 2020, about the same proportion as live in slum conditions. Given the impact of safely managed sanitation on health and mortality, the need for improved sewer access is urgent, and improving such access is one of the United Nations' "Millenium Development Goals." Yet, many cities also lack decent roads, sufficient public transit, adequate schools, and reliable electricity. Trade-offs between these services must be evaluated and made. Our finding that improved sewer access causes economically large increases in density but does not precipitate the arrival of more affluent migrants, should be of immediate use to policy makers evaluating such trade-offs.

Urban migration is among the best known ways to increase individual wages in developing countries (Gibson et al., 2014, Lagakos et al., 2020). Henderson and Turner (2020) estimate that for a typical resident of the developing world, moving to a location that is twice as dense increases household incomes by 32%. This leads us to ask why developing world countries are not urbanizing faster. One possibility is that developing world cities are difficult places to live, in part because they often lack basic sanitation. Our results strongly support this conclusion. Our results suggest that, by facilitating increased population density, improved sewer access can allow cities to accommodate more of the the rural poor. Indeed, by providing estimations of the magnitude of the effect of sewers on population density, we are providing a foundation for the cost-benefit analysis of sewer expansions that includes benefits to rural migrants.

Despite its importance, the effect of sewer access on urban development has received little attention from researchers. There appear to be two reasons for this. First is the difficulty in organizing systematic descriptions of sewer networks. Sewers are underground, often old, and often administered locally, all factors that increase the difficulty of data collection. Second is the fact that sewers are not assigned to places at random, and the literature has failed to develop a quasi-experimental research design to address this problem that can be widely applied. We solve both problems. We exploit GIS technology to develop a quasi-experimental design using widely available census data and universally available digital elevation maps.

2 Literature

There is a large literature studying the effects of urban infrastructure. For example, Jedwab and Storeygard (2022) and Ghani et al. (2016) study the effects of highways and roads in India and Africa; Tsivanidis (2019) studies the effects of bus rapid transit in Bogota; Gendron-Carrier et al. (2022) studies the effects of subways all over the world; and finally, Allcott et al. (2016) and Lipscomb et al. (2013) study the effects of electrification in India and Brazil.

There is also a literature studying the effect of water quality on health outcomes, usually infant and child mortality, in the developing world (e.g., Ashraf et al. (2017), Galiani et al. (2005), Bhalotra et al. (2021)) and in the developed world during the industrial revolution (e.g., Anderson et al. (2018), Ferrie and Troesken (2008), Kesztenbaum and Rosenthal (2017), Ogasawara and Matsushita (2018)). These studies usually find large effects of improved water quality on health and mortality.

Studies of sewers are rarer. Alsan and Goldin (2019) study late 19th century Boston and find a large reductions in infant mortality from the joint roll-out of municipal water and sewer systems, but no evidence that people sorted into places with better water and sewer service on the basis of observable demographics. Anderson et al. (2018) examine the effect of sewer system construction in 25 US cities in the early 1900s and, contrary to Alsan and Goldin (2019), find no relationship between measures of mortality and sewage treatment or the interaction of sewage treatment and water treatment.

To our knowledge, Gamper-Rabindran et al. (2010) is the only paper to

explicitly study urban sewer systems in the developing world. This paper considers a municipality-year panel of Brazilian data reporting infant mortality and municipal level measures of water and sewer access. They find that access to piped water, but not to sewers, has a large effect on infant mortality. Only Coury et al. (2022) explicitly considers the relationship between sewer construction and urban development. Coury et al. (2022) investigates the effect of expansions of the Chicago water and sewer network in the late 19th century on the price of residential land. They find that sewer and water access more than doubles land prices.

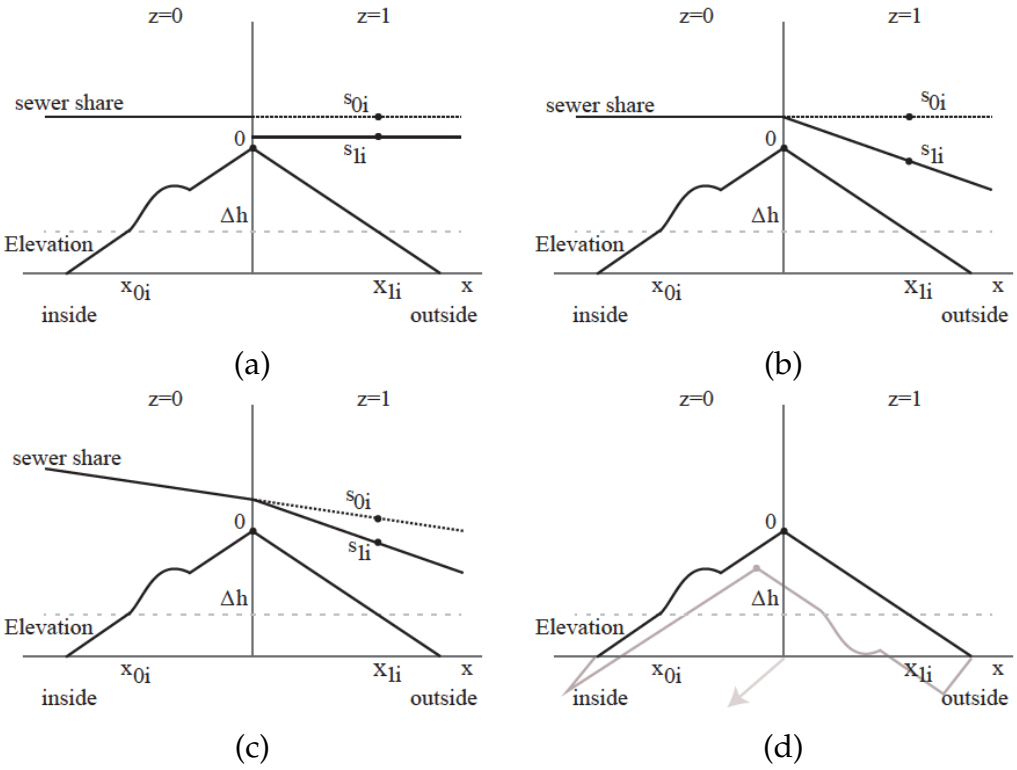
Summing up, the available evidence suggests that sewer access has beneficial effects on cities and neighborhoods. The evidence for more specific effects is thin and based on 19th century US cities. Coury et al. (2022) find that sewer access leads to large increases in land prices in late 19th century Chicago. It is natural to suspect that these prices increases were associated with an increase in density. Extrapolating from the Alsan and Goldin (2019) finding that people do not sort on the basis of local variation in municipal water and sewer service suggests that people should not sort on the basis of sewer access. However, there is little evidence about the magnitudes of these effects in contemporary developing world cities, or about heterogeneity in these effects across places. These are the questions we address.

3 Identification

The movement of wastewater in sewers is special in two regards. First, it is sensitive to variation in elevation that is irrelevant for most other activities. Gravity sewers require a grade of about 1:200 (1 unit of drop per 200 of horizontal). While the details of pipe size, shape, and interior smoothness can partially compensate for vertical drop, in general, at a grade of about 1:200 solids settle out of the flow and block the pipe (Mara, 1996). For reference, athletes will generally perceive a playing field as sloped only once it has a grade of more than 1:70 (Aldous, 1999).

Second, unlike people, wastewater only travels away from a residence. Thus, commuting should respond symmetrically to elevation change on the outbound and inbound trips, but sewers should respond asymmetrically. For household wastewater, only elevation gain outbound is costly.

Figure 1: Identifying treatment effects around a stylized basin boundary



Note: *Elevation and sewer share profile in the neighborhood of a drainage basin divide. The basin divide is at the top of the hill, at $x = 0$. Displacement left is "inside" and towards the nearest established sewer system. Displacement right is "outside" and wastewater in this region must travel uphill to reach the nearest sewer network. (a) Crossing the basin divide is a discrete shock to the cost of sewer access. (b) Crossing the divide increases the cost of sewer access continuously with distance to divide. (c) Same as (b) but x displacement has an independent effect on sewer access. (d) Illustration of variation in elevation independent of x.*

These two facts motivate the identification strategy illustrated in figure 1. The peaked dark line in this figure describes the elevation profile along an axis horizontal to a drainage basin divide at $x = 0$. The region to the left of $x = 0$ is "inside" the central city drainage basin and drains downhill to the sewer system servicing the CBD. The region to the right of $x = 0$ is "outside" and cannot reach the central city sewer network without travelling uphill.

Moving sewage across a basin divide is difficult and may be accomplished in three ways (Mara, 1996). First, by burying sewer pipes more deeply, the grade of the sewer pipe can diverge from that of the ground above. Recalling that a sewer needs a grade of 1:200, burying a sewer to a depth of eight feet instead of two

can allow an extra 1200 feet of horizontal travel. Second, if the topography allows, following an indirect route approximately along an elevation contour to reach the inside of the basin allows the substitution of downhill, horizontal travel for climbing. Third, building pumping facilities to lift the sewage over the basin divide is also possible. This requires the availability of electric or fossil fuel powered pumps. If the land outside the central basin is sufficiently valuable, there is also the possibility of building a new sewer network to serve the relevant drainage basin and to avoid moving sewage up and across a basin divide altogether. All four possibilities are costly. Crossing a drainage basin divide from a basin with sewer service to one without increases the cost of sewer access.

Summing up, for places on the outside of a drainage basin divide, the cost of reaching the central city sewer network should increase rapidly with the horizontal and vertical distance that sewage must cover to reach the basin divide (from which it can drain downhill to the central city sewer network). Conversely, for places on the inside of the basin divide, horizontal and vertical displacement from the divide should have less impact on the cost of sewer access, or none at all.

As we will see, drainage basin divides are usually almost unnoticeable landscape features. From this it follows that locations close to, but on opposite sides of a drainage basin divide should be similar in their suitability for urban use, except that sewers will be more costly for outside locations. This suggests that for a sample of locations close to a drainage basin divide, being inside or outside the basin is a source of quasi-random variation in the cost of sewers. Our research design is organized around comparing census tract level sewer access and demographics in nearby tracts on opposite sides of a drainage basin divide.

The four panels of figure 1 inform the exercise of translating this intuition into an econometric specification. The horizontal axis, x, is displacement along a horizontal axis perpendicular to a drainage basin divide, henceforth simply "horizontal displacement." In all four panels, locations to the left of zero are inside and are uphill from the sewer system serving the central city. Locations to the right are outside, and sewage in these locations must travel horizontally and vertically to the basin divide before it can drain to the center.

Define a binary variable $\mathbb{1}(\text{Outside})(x)$ equal to one for x outside, and zero otherwise. Let s indicate the share of households in a location with sewer access,

and define $\Delta h(x) \geq 0$ as meters of descent required to reach x from the top of the basin divide at $x = 0$. Thus, $\mathbb{1}(\text{Outside})(x)x$ and $\mathbb{1}(\text{Outside})(x)\Delta h(x)$ are the horizontal and vertical displacement required to reach the inside of the central drainage basin from location x. We consider $\mathbb{1}(\text{Outside})(x)$, $\mathbb{1}(\text{Outside})(x)x$, and $\mathbb{1}(\text{Outside})(x)\Delta h(x)$ as instruments.

Our measure of sewer access is the share of households in a census tract reporting that they have access to a public sewer. The size of census tracts varies by country, but they are usually at least one half kilometer square. At this scale, it is possible that the cost shock to sewer construction will appear instantaneous when we cross the basin divide. This case is illustrated in panel (a). In this figure, we suppose that sewer share, s, does not depend on x, except at the basin divide, where the cost of sewer access increases, and the share of houses reporting access to a sewer declines as a step function.

We would like to estimate how an outcome Y depends on sewer access. If panel (a) is an accurate description of the world, then we can do this by comparing the size of the step down in sewer access at $x = 0$ to the corresponding change in Y.

However, it is hard to have a strong prior about the spatial scale over which the basin divide cost shock will operate and there are reasons to think that it will not operate as sharply as illustrated in panel (a). The area near a drainage basin divide is often quite flat. Moving a few hundred horizontal feet outside of the basin divide may involve only a foot or two of drop. Because each foot of vertical drop allows about 200 feet of horizontal travel, in flat terrain, burying a sewer eight feet deep rather than two allows about an extra 1200' of access. Moreover, we probably measure the locations of basin divides imprecisely, so measurement error should smooth out the empirical counterpart of figure 1(a). In either case, we expect sewer access to decline smoothly with distance to the basin divide.

The case when sewer access declines continuously as we move progressively further outside the the central basin is illustrated in figure 1(b). The intuition behind panels (a) and (b) is similar, but the implied econometric model is not. Panel (a) can be described by a discrete instrument, and a discrete treatment. In panel (b) the cost shock increases in distance, as does the resulting change in sewer share, so instrument and treatment are both continuous. The econometrics of estimating treatment effects with continuous and binary treatments are quite

different.

Figure 1(c) illustrates one of the main challenges to our identification strategy. By construction, each central basin encircles the center of the city it contains. Thus, displacement inside is usually towards the city center and conversely. As we move towards the center, we expect land to become more valuable and more intensively developed. Assuming the cost shock to sewers operates continuously, as illustrated in figure 1(c), we expect a steady decline in sewer share as we move from left to right, away from the city center, with a trend break and more rapid decline once we cross the basin divide.

Figure 1(d) illustrates a final point about our identification strategy and suggests a different instrumental variable. In reality, and unlike what we illustrate in the first three panels, our data will lie on a strip rather than a line. This means that there will be variation in elevation, holding distance to the basin divide constant. Therefore, we can estimate the effect of elevation on sewer access, conditional on x. Holding x constant, we expect vertical displacement to have a larger effect on sewer access outside the basin divide than inside.

4 Central Basins

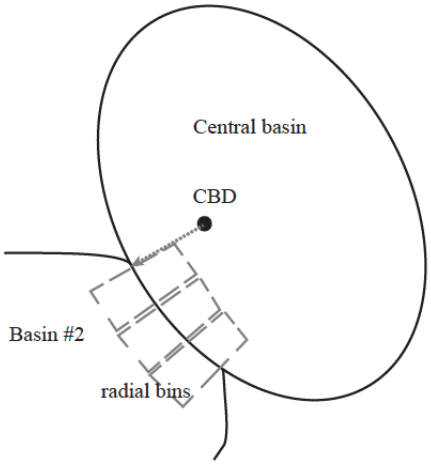
We now turn to the problem of defining an empirical analog to the illustrations in figure 1, and to defining "inside" and "outside."

Figure 2 describes the intuition behind our approach. The central ellipse in this figure describes the drainage basin containing a central city with a sewer system. This is the "central basin." All points in this basin drain to the same point, and so, in principal, can be served by the same sewer network. Any point not in the central basin, by construction, does not drain to this point. The boundary of the central basin is the central basin divide and a location is "inside" or "outside" as it lies inside or outside the central basin.

To construct central basins, we begin with the UN DESA World Urbanization Prospects data. These data report the coordinates of the centers of all cities that have a population of 300,000 or above in 2018 (UN DESA Population Division, 2018). We restrict attention to the cities in countries where we have census data and maps: Brazil, South Africa, Tanzania, Jordan, and Colombia.

We next download the Advanced Spaceborne Thermal Emission and

Figure 2: Illustration of basins, segments, radial-bins, and "inside" indicator



Note: *The central ellipse describes the drainage basin containing a center city. The boundary of this drainage basin is the central basin divide. A location is "inside" or "outside" as it lies inside or outside the central basin. The central basin generally abuts other drainage basins. The portion of the central basin divide which divides a particular pair of basins is a "segment" of the basin divide. We divide the area near the basin divide into "radial-bins"(sometimes "π-bins"). To construct these bins, we divide the central basin divide into two kilometer long intervals, starting from the point on the basin divide nearest the city center. A radial-bin is the area within* 2KM *of one such interval.*

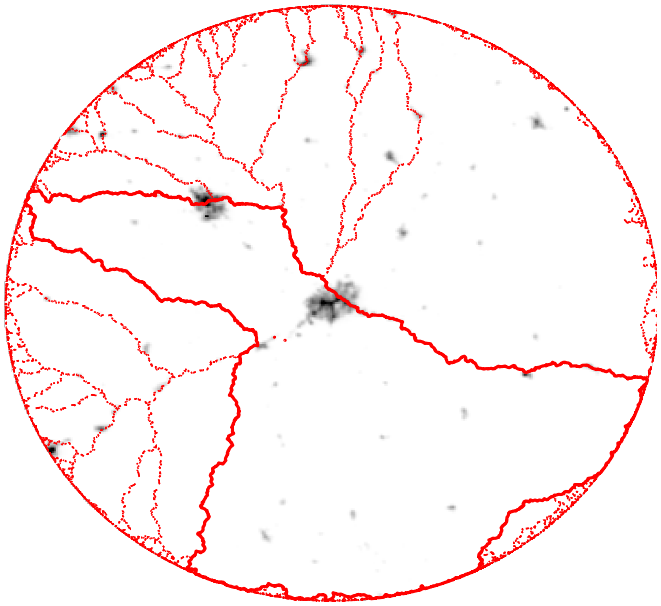
Reflection Radiometer (ASTER) (NASA/METI/AIST/Japan Spacesystems and US/Japan ASTER Science Team, 2019) digital elevation map. These data report the elevation of most of the Earth's surface at a spatial resolution of about $30m^2$.

From the ASTER data, we clip out a circle of radius 75KM centered on the CBD of each sample city. Each such circle is an elevation map of one of our cities and its hinterlands. This done, we draw all drainage basins within a 75KM radius of the center of each city using an ArcGIS utility. Finally, we identify the drainage basin containing the center of each city. These are the central basins, and their boundaries are the central basin divides.

We will ultimately rely on census data at the tract level to describe sewer access and outcome variables, and our unit of observation will be a census tract. We include in our sample only census tracts for which the closest basin divide is the central basin divide for one of the cities in the UN DESA data. Our research design is organized around comparisons of census tracts on opposite sides of central basin divides.

We define a census tract as inside or outside depending on whether its area

Figure 3: Drainage basins containing Cascavel, Brazil



Note: *Dashed Red lines indicate drainage basins boundaries based on the* ASTER *digital elevation map. The solid red line indicates the basin boundary for the basin containing Cascavel, Brazil.* VIIRS *lights at night shows city extent. The disk has a radius of 75*KM.

weighted centroid is on the same side of the closest basin divide as the central business district contained in all of our central basins. To be included in the UN data a city must have a population of at least 300k. As a result, all central cities in our sample have at least some sewer service.[1] Therefore, this definition guarantees that an inside census tract can drain to a central city sewer network.

Figure 3 illustrates basin boundaries around Cascavel, Brazil, and is an empirical analog of figure 2. Red dashed lines indicate the boundaries of all drainage basins, and solid red shows the boundary of the central basin. Shading is based on lights at night and shows the scale of the city relative to the various basins.

In figures 2 and 3, each central basin abuts other drainage basins. The portion of the central basin divide which divides a particular pair of basins is a "segment" of the basin divide. For econometric purposes discussed later, we

[1]We experimented with using lights-at-night weighted tract centroids and found that it did not have an important effect on our results.

divide each central basin divide into these segments. We also divide the area near the basin divide into "radial-bins," which we sometimes abbreviate to "π-bins." To construct these bins, we divide the central basin divide into two kilometer long intervals, starting from the point on the basin divide nearest the city center. A radial-bin consists of all census tracts with centroids within 2KM of one such 2KM interval. We define "segment-bins" analogously on the basis of basin segments rather than two kilometer intervals.

In our preferred regression specification below, no radial-bin containing fewer than three census tract centroids can contribute to the identification of the causal effect of sewer access. In order to assure that less restrictive specifications and descriptive statistics describe the same variation in the data, we generally drop any tract that lies in a radial-bin containing fewer than three tract centroids.

Because two central basins may be adjacent, our notion of inside and outside can be undefined. There are two natural solutions to this problem. The first is to exclude all such tracts from our sample. Alternatively, for tracts for which the closest basin divide segment divides two central basins, define inside and outside on the basis of the closest of the two city centers. We experiment with both strategies and our results are robust to either definition. However, when this situation arises, on average, the more remote CBD is three times as far away as the closer one. Therefore, while this notion of inside and outside can be ambiguous in theory, it is rarely ambiguous in practice. Given this, we report results based on the larger sample.

Looking carefully at figure 3, we see that our basin drawing algorithm sometimes constructs incoherent basins at the edge of the map disc. For this reason, we exclude from our study the region within 6KM of the edge of these discs, or conversely, more than 69KM from the center of the city.

The drainage basins that contain the coordinate of the CBD for our cities are sometimes too small to contain a meaningful share of the city's population (recall the UN DESA data reports on cities with a population above 300,000). This is a particular problem for cities near the coast, where the basin drawing algorithm tends to construct small basins. To see why this creates a problem consider two central basin segments, one about 100 meters from the CBD, and one 10KM from the CBD. For the first, displacement inside the basin divide is displacement towards the CBD for about the first 100 meters, and then it is displacement

beyond and away from the CBD. In the second case, displacement away from and inside the basin divide is towards the CBD for about 10KM. Pooling these two types of basin divides complicates the interpretation of horizontal displacement. To resolve these problems, we define our central basins as the union of drainage basins that intersect a disk of 2KM radius centered on the CBD. This guarantees that no point on the central basin divide is closer to the CBD than 2KM.

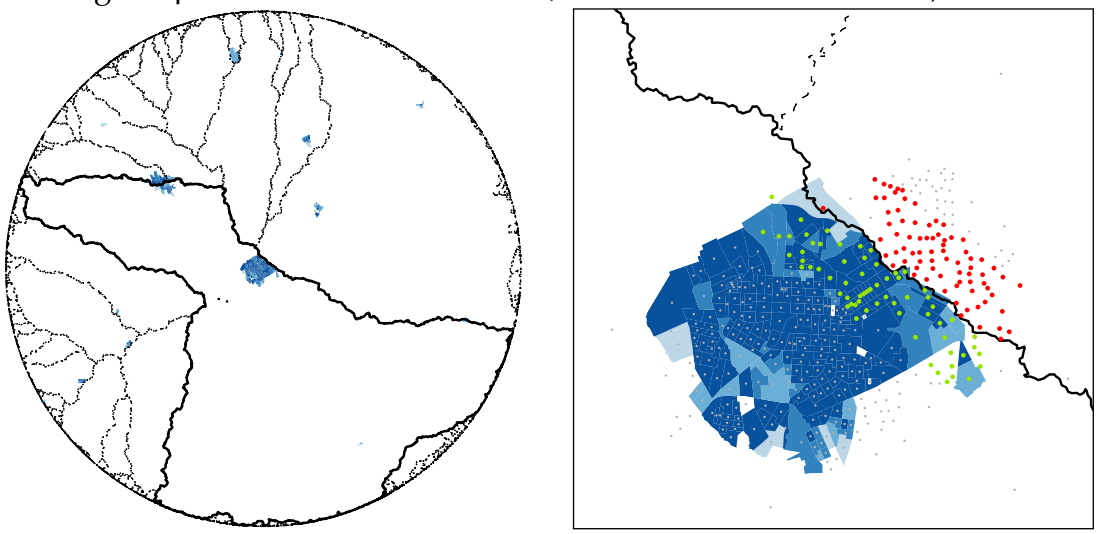
A problem may arise when small towns lie close to, but outside a central basin divide, and are far from the main CBD. If these small towns have sewer networks, then for census tracts close to these small towns, being inside the central basin probably places them farther from the nearest sewer network. We experimented with alternative definitions of inside that address this problem. These alternatives face the following problem. We can only measure sewer access with the same census data that we use to define our treatment. Thus, picking out small, highly sewered towns, near the basin divide relies on the same data we use to construct our treatment variable. Therefore, any definition of inside based on these data implicitly requires that we condition on an endogenous variable. Given this, we do not pursue alternative definitions of inside and outside.

5 Sewers and outcomes

The Brazilian census asks households if they have a toilet, and whether this toilet drains to a sewer, septic tank, a ditch, a pit, or surface water. In this way, the census provides an indicator of "connected to a sewer." These data are publicly available, aggregated to the "sector" (about the same size as a US census block group). Equivalent questions appear in the census forms of Colombia, Tanzania, Jordan, and South Africa. Appendix B provides more detail about these census questions. These questions allow us to calculate the share of households in a census tract with access to a sewer for our entire sample of tracts. This is our treatment variable, and we would like to know the effects of changes in the share of tract households with sewer access.

It remains to calculate the vertical distance between each tract and the central basin divide. For this purpose, we calculate the height of the basin divide for each tract as the highest elevation of any tract centroid in the same radial-bin as the target tract. We then calculate the vertical rise required to reach the basin divide, $\Delta h(x)$, as the elevation difference between the target tract centroid and

Figure 4: Sewer share and "inside/outside" around Cascavel, Brazil



Note: *Full basin and close up of Cascavel, Brazil. Darker blue indicates a larger share of households reporting a toilet connected to a public sewer. Dots indicate census tract centroids. Centroids for which the closest basin divide is not the central basin are excluded (light gray), as are centroids that are more than* 2KM *from the central basin divide. "Inside" centroids are green, and "outside" centroids are red. Because basin boundaries are often incoherent near the edge of the* 75KM *disk of the* DEM *we work with, tracts with centroids more than* 69KM *from the city center are also excluded from the sample.*

this radial-bin maximum.

Table A1 provides a country-by-country breakout of our data. Three features of this table are noteworthy. First, Brazil accounts for the largest share of cities in our data by a wide margin, but the number of Colombian and Brazilian tracts in our data is about the same. South Africa has about the same number of cities as Colombia, but about a quarter the tracts. Jordan accounts for a small share of cities and a smaller share of tracts. Second, the average over tracts of sewer share is about 0.7 for Brazil and Columbia. It is about 0.8 for South Africa, 0.6 for Jordan and about 0.08 for Tanzania.

Having defined our treatment as "share of tract households with sewer access," it is of interest to know what is the alternative to sewer access. Colombia's census question is binary. A household reports sewer access or not. However, Brazil, Jordan, South Africa, and Tanzania each provide more detail, although it is hard to compare "not sewered" outcomes across countries. Table A4 organizes these data. The two main alternatives to sewer access are cesspits

and septic tanks, respectively a hole in the ground (possibly lined) and a lined tank in the ground. A third category consolidates "other" and "none." For the four countries where we can refine the not sewered category, our estimation sample describes about 5.8m households with an average of 3.3 people per household. Of these, about 68% have sewer access, about 28% have a cesspit or septic tank. The remainder have no sanitation facilities, or some arrangement other than a cesspit or septic tank.

Septic systems are fairly common in low density development in the US. These systems consist of a septic tank and a large, highly regulated drain field. For example, the zoning code of Lawrence County South Dakota prohibits septic systems on lots of less than two acres.[2] The high population densities observed in our census tracts means that the septic tanks that prevail in our sample must be quite different from those in the rural US. In all, this suggests that sanitation for the unsewered households in our sample is primitive.

Figure 4 is a heat map illustrating the incidence of sewer access for the Brazilian city of Cascavel. Polygons describe the extent of census tracts, with darker blue indicating a larger share of households reporting sewer access. Basin boundaries are black lines. Dots indicate census centroids. Census centroids are red if they are inside the central basin, green outside, and gray if excluded from our sample.

The Brazilian, Columbian, South African and Tanzanian censuses report population by tract. Because we have GIS maps of tract boundaries, we can also calculate tract area, and hence tract population density. Jordan's census reports only the count of households, so for Jordan we use household density in place of population density. The Brazilian and South African censuses report on household income. Appendix B describes how we construct our income variable from the available census questions for these two countries. Each of the censuses for Tanzania, South Africa, Brazil and Colombia reports some information on educational attainment or literacy. We use these questions to create a standardized measure of the share literate in each tract in these four countries (see Appendix B for details).

Summing up, our data describes census tract equivalent units in Brazil,

[2] https://codelibrary.amlegal.com/codes/lawrencecounty/latest/lawrencecty_sd_land/0-0-0-2364 accessed February 14, 2025.

Colombia, South Africa, Tanzania and Jordan. We restrict attention to tracts whose centroids are (1) within 2KM of the nearest central basin divide, (2) within 69KM of the city center, and fall in a radial-bin containing three or more tract centroids. We assign sewer share, population density and other outcomes on the basis of the relevant census survey data, and vertical distance to basin divide on the basis of the elevation of the highest tract centroid in the relevant radial-bin.

Table 1 describes the sample on which we base our estimations. Column 1 reports on all tracts with centroids that fall in the central basin of one of the 92 cities in our sample. This column describes the cities of interest. This sample consists of about 240,000 tracts and 4,000 radial-bins. Column 2 describes all tracts that (1) lie within 2KM, inside or outside, of the central basin divide for one of these 92 cities, and (2) lie in a radial-bin containing at least three tract centroids. These are the tracts we will use to estimate the effects of sewer access. This sample consists of about 50,000 tracts and about 1,500 radial-bins. We refer to the sample described in column 1 as the "cities sample" and to the sample described in column 2 as the "estimation sample".

As expected, the estimation sample is about 1KM from the basin divide on average and is about evenly split between inside and outside tracts. On average, a tract in the estimation sample is about 12KM from the CBD. Other results are more surprising.

By construction, tracts in the estimation sample are further from the CBD than those in the cities sample. In spite of this, mean distance to the CBD is larger in the cities sample. There are two reasons for this. First, the table presents tract weighted averages and cities located in large drainage basins have more tracts that are further away. As a result, restricting to tracts within 2KM of the basin divide drops many more tracts in large basin cities than in small basin cities, reducing the average distance to CBD when pooled across cities.[3] Second, the estimation sample excludes radial-bins containing fewer than three tract centroids. This excludes more remote and less dense locations included in the cities sample. Population densities are high in both samples. It is tempting to think that this reflects city wide density. This is not correct. Because dense areas

[3]Within each city, the sample restriction does increase the average distance to the CBD as anticipated.

contain many small tracts, dense areas are overweighted in these tract averages.[4] The final two rows of the table report income per month and share literate. As described above, the income data reflects only Brazil and South Africa, and the share literate excludes tracts in Jordan.

6 Descriptive results

Figure 5 shows empirical analogs of the elevation profile in figure 1. Each of the three panels presents a binscatter plot of the mean bin elevation as a function of the distance from each tract centroid to the nearest point on the central basin divide. As in figure 1, the basin divide is at $x = 0$, left of zero is inside the central basin and right outside. All three panels are based on the estimation sample described in column 2 of table 1.

The top panel shows unconditional bin means. In this figure, the expected high point at the basin divide is on average lower than interior tracts. In the middle panel, we repeat the exercise using tract centroid elevations net of city mean elevation. In this panel, we begin to see the expected high point at the basin divide. In the bottom panel, we repeat the exercise again, but based on tract centroid elevation net of radial-bin mean elevation. We see the expected peak at the basin divide clearly.

This figure demonstrates three important features of our data. First, the drainage basin divides are not dramatic geological features in two senses. First, comparing the bottom panel of figure 5 to the panels above, we see that the variation in elevation associated with a few KM of travel perpendicular to the basin divide is small relative to the variation across cities, or the variation within a city as we travel circumferentially along the basin divide. Second, the bottom panel of figure 5 also shows that the basin divides are small features in an absolute sense. On average, traveling 2KM horizontally to the basin divide, whether inside or outside, involves a descent of about 30m. Thus, the average grade along a 2KM path extending from 2KM inside the basin divide to 2KM outside is about 1:70, the grade at which athletes begin to notice a playing field

[4] To see the importance of this, consider a city with just two tracts: the first tract has 250 people and geographic area $= 0.01$ KM^2, and the second tract has 250 people and area 2KM^2. Then the average population density for the whole city is $(250 + 250)/(2 + 0.1) \approx 249$ people per KM^2 but the average of the tract level densities is more than 12,500 people per KM^2 ($\frac{1}{2}(\frac{250}{0.01} + \frac{250}{2}) \approx 12{,}563$).

Table 1: Descriptive statistics

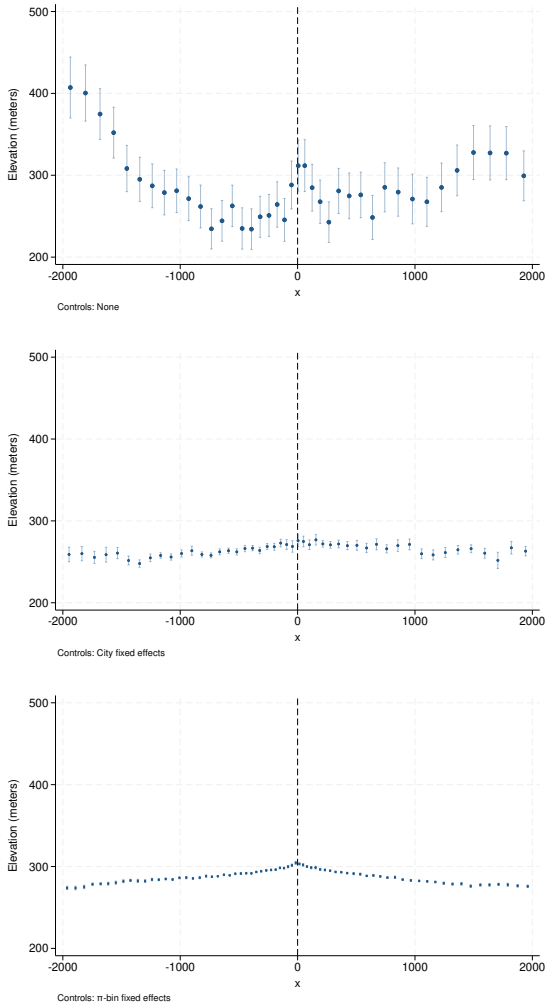
	(1) CBD Basin + 2KM	(2) ± 2KM Basin Divide
Cities	92	92
Mean area CBD basin (KM^2)	1,371	1,371
Segments	822	500
π-bins	4,094	1,513
Tracts	239,393	50,039
Share inside	0.90	0.54
	(0.30)	(0.50)
Mean tract area (KM^2)	0.63	0.16
	(5.80)	(1.25)
Mean dist to CBD (KM)	13.53	12.06
	(15.04)	(16.10)
Mean log dist to CBD (m)	8.95	8.71
	(1.11)	(1.15)
Mean dist to basin divide (KM)	11.02	0.86
	(8.93)	(0.59)
Sewer share	0.75	0.69
	(0.33)	(0.37)
Mean num people in a tract	339	427
	(367)	(392)
Pop density (persons/KM^2)	28,379	22,348
	(34,882)	(27,093)
Income (per month, 2022 USD)	968	940
	(921)	(834)
Share literate	0.93	0.92
	(0.30)	(0.40)
Elevation (m)	939	288
	(799)	(520)

Note: *Column 1 describes all tracts with centroids falling in one of the 92 central basins that make up our sample, or less than 2KM outside. All cities lie in Brazil, Colombia, Jordan, South Africa, or Tanzania. Column 2 describes our main estimation sample. It consists of all tracts that (1) fall within 2KM of the central basin divide for one of the 92 cities in our sample, and (2) fall in a radial-bin containing at least three census tract centroids. Tract mean income is based only on tracts in Brazil and South Africa. Tract literate share data excludes Jordan.*

is sloped.

Second, the peak at $x = 0$ visible in the bottom panel of figure 5 is by construction. Basin divides are constructed to lie at local high points. That we cannot see the basin divide in the top two panels indicates that when we are comparing tracts across the basin divides, we are not making the comparisons we intend. It is only once we control for radial-bin means that we seem to be

Figure 5: Mean tract centroid elevation and conditional elevation as a function of distance to the nearest basin divide

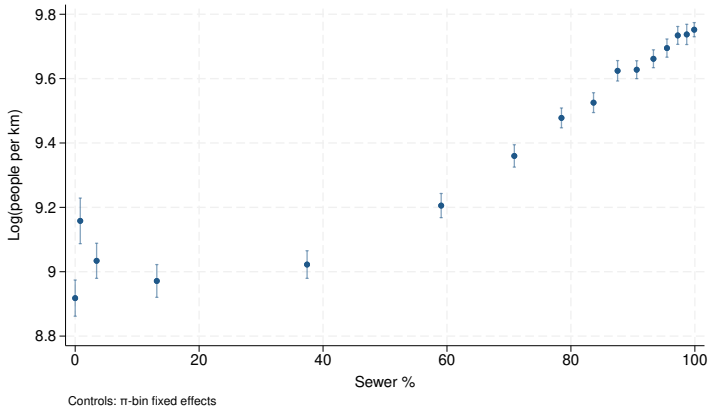


Note: *Mean elevation by distance to basin divide; raw data (top), net of city mean (middle), and net of radial-bin mean (bottom). On average the divide is not a dramatic feature. Figures based on the estimation sample described in column 2 of table 1.*

comparing tracts that are "close enough" together that the expected pattern in the data emerges. This motivates our reliance on radial-bin level variation in our estimations.

Third, whiskers in the bottom panel describe variation in elevation around the mean, conditional on displacement from the basin divide. This is variation in elevation holding horizontal displacement constant, exactly the variation that our second instrument exploits. The confidence intervals around the trend line

Figure 6: Logarithm of tract population density vs sewer share



Note: *Mean log population by tract sewer share. Log tract population density increases from about 8.8 to about 9.8 as sewer share increases from zero to one. On average, each 1 percentage point increase in sewer share is associated with about a 1% increase in population density. Figure based on the estimation sample described in column 2 of table 1.*

in the bottom panel of figure 5 give a sense for the magnitude of this variation. These tight confidence intervals also indicate that most basin divides are nearly unnoticeable features of the landscape. There are not many basin divides that are marked by dramatic changes in elevation.

Figure 6 shows the correlation between the share of households in a tract with sewer access and the logarithm of tract population density. The figure is a binscatter plot, and so the slope reflects means in the raw data. Throughout most of the range of sewer access, the relationship is approximately linear, and the slope indicates an elasticity around one. That is, each 1 percentage point increase in sewer access is associated with a 1% increase in population density. Because we expect that the assignment of sewer access to census tracts is not done at random, we cannot interpret this slope as a causal effect of sewer access on density. Estimating this casual relationship is the central econometric problem that we address.

Figure 7 illustrates the empirical variation relevant to our first identification strategy. All three panels are binscatter plots reporting means for a different variable as a function of distance to a central basin divide, net of radial-bin means.

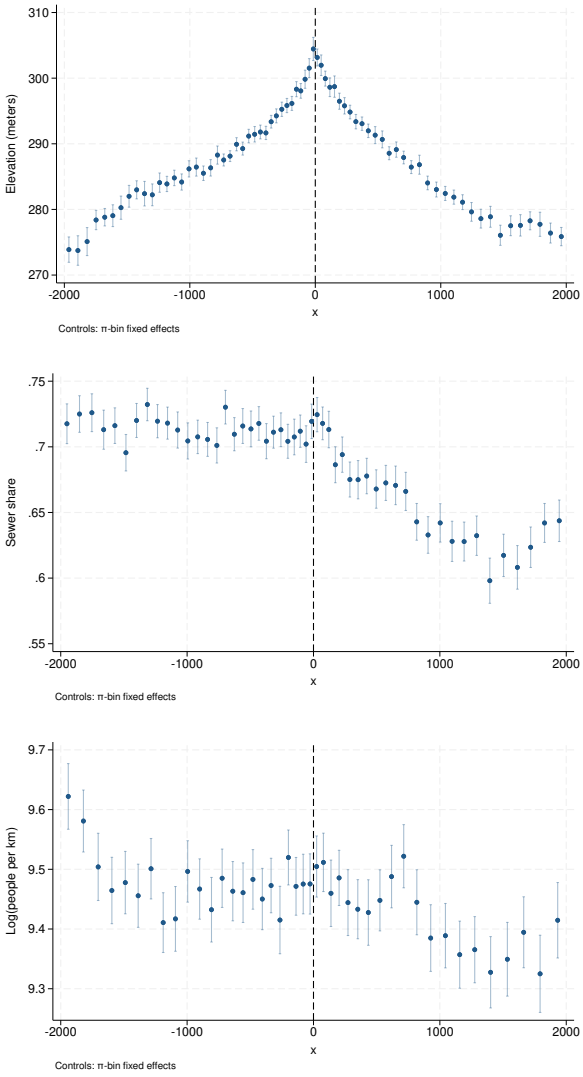
For reference, the top panel repeats the bottom panel of figure 5, but with the y-axis rescaled. Traveling away from a central basin divide means traveling downhill, slightly. The middle panel shows changes in sewer share as a function of distance to the central basin divide, net of radial basin means. This is a first-stage regression. We see a trend break in sewer share at $x = 0$, and possibly a small step down. This is consistent with our intuition about the costs of sewer construction. Crossing the basin divide increases the cost of sewers and decreases their prevalence, but the exact functional form of this relationship is not obvious, and may be confounded by an independent effect of elevation or horizontal displacement.

The bottom panel of figure 7 is like the middle panel, but reports bin means of log tract population density. This is a reduced form regression. This figure is less clear than the corresponding figure for sewer share, but population density appears to decline outside the central basin.

Our econometric specification exploits the variation illustrated in figure 7 by including an indicator for whether a tract is outside and the interaction of this indicator with horizontal distance to the basin divide. The validity of these instruments probably depends on controlling for elevation and horizontal displacement, and we experiment with different specifications using these controls.

Figure 8 illustrates the empirical variation relevant to our second identification strategy. Like figure 7, both panels in figure 8 are binscatter plots. Like the middle and bottom panels of figure 7, the y-axis of the top and bottom panel of figure 8 reports bin means of sewer share and log tract population density net of radial-bin means. However, the x-axis of these figures is different from figure 7. In figure 8 the x-axis describes meters of climbing required to reach the central basin divide. Positive x values indicate meters of climbing to reach the basin divide for a tract on the outside of the basin, and negative x values indicate meters of climbing to reach the basin divide for a tract on the inside of the basin. For example, the mean sewer share is about 0.68 for tracts that are 100m below and outside the basin divide, while the mean sewer share is about 0.75 for tracts that are inside the basin divide and only a few meters below it. This figure is also a first-stage regression. Looking carefully shows a clear break in sewer share at $x = 0$. Mean sewer share is about 0.08 lower for a tract a

Figure 7: Identification, $z = \mathbb{1}(x \text{ is outside})$

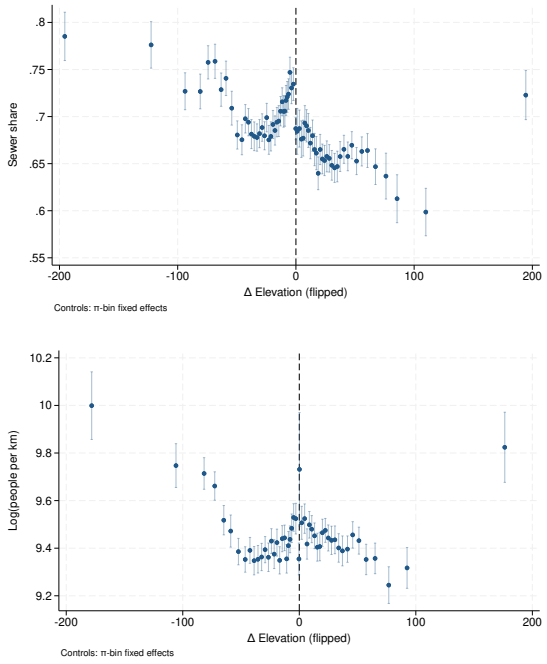


Note: *All panels are binscatter plots. (Top) Mean tract elevation net of radial-bin fixed effects as a function of horizontal displacement. (Middle) Mean tract sewer share net of radial-bin fixed effects as a function of horizontal displacement. (Bottom) Mean tract population density net of radial-bin fixed effects as a function of horizontal displacement. Top and middle are the empirical analogs to figure 1. All three panels are based on the main estimation sample described in column 2 of table 1.*

few meters below and outside the basin divide than for a tract just below and inside the divide.

The bottom panel of figure 8 is the corresponding reduced form. This

Figure 8: Identification, $z = \mathbb{1}(x \text{ is outside}) \times \Delta h$



Note: *Both panels are binscatter plots. Top panel reports mean tract sewer share as a function of vertical distance to basin divide, net of radial-bin mean. Bottom panel is mean log tract population density as a function of vertical distance to basin divide, net of radial-bin mean. Both panels are based on the main estimation sample described in column 2 of table 1.*

binscatter plot is constructed in the same way as the top panel, but the y-axis reports the bin mean of log tract population density. This figure also shows a trend break and a step at $x = 0$.

Our econometric specification exploits the variation illustrated in figure 8 by including an indicator for whether a tract is outside and the interaction of this indicator with vertical distance to the basin divide, while controlling for horizontal displacement and elevation.

7 Reduced form results

We begin by estimating the effect of the tract share of sewer access on tract population density. To proceed, let j index census tracts and k index radial-bins. s_{jk} is the share, from zero to one, of households reporting sewer access in tract j and radial-bin k. y_{jk} is the outcome of interest, the logarithm of population density. x_{jk} is meters from the tract centroid to basin divide, with displacements

inside the basin negative and displacements outside positive. $\Delta h_{jk} \geq 0$ is the vertical rise required to reach the basin divide from the centroid of tract j and $\mathbb{1}(\pi\text{-bin})_{jk}$ is an indicator that is one for all tracts in radial-bin k and zero otherwise. Finally, let $\mathbb{1}(\text{Outside})_{jk}$ be an indicator variable that is one for tracts with centroids outside the central basin, and zero otherwise.

Our research design requires two estimating equations. The first is a first-stage predicting sewer share using all three (or a subset) of our instruments, $\mathbb{1}(\text{Outside})_{jk}$, $\mathbb{1}(\text{Outside})_{jk}x_{jk}$, and $\mathbb{1}(\text{Outside})_{jk}\Delta h_{jk}$, along with controls for radial bin and elevation,

$$\begin{aligned} s_{jk} =& \mathbb{1}(\pi\text{-bin})_{jk} + \mathbb{1}(\pi\text{-bin})_{jk}x_{jk} + A^s\Delta h_{jk} + \\ & \alpha_0\mathbb{1}(\text{Outside})_{jk} + \alpha_1\mathbb{1}(\text{Outside})_{jk}x_{jk} + \alpha_2\mathbb{1}(\text{Outside})_{jk}\Delta h_{jk} + \eta^s_{jk}. \end{aligned} \tag{1}$$

The second is a structural equation predicting a tract outcome as a function of tract sewer share and controls,

$$y_{jk} = \mathbb{1}(\pi\text{-bin})_{jk} + \mathbb{1}(\pi\text{-bin})_{jk}x_{jk} + A\Delta h_{jk} + \beta s_{jk} + \eta_{jk}. \tag{2}$$

Depending on estimation technique, equation (2) is an OLS estimation or a TSLS/IV regression.

These equations require several comments. First, at the parcel level, sewage access is binary. A parcel has access or it does not. In our tract level data, we observe the share of parcels treated. If the parcel-level causal effects are homogeneous then treatment effects are also homogeneous at the tract level. In this case we can interpret the coefficient of the sewage share in (2) as reflecting the parcel level causal effect of sewer access.

Second, figure 5 demonstrates that the expected elevation profile around basin divides is only present once we control for radial-bin fixed effects. Because of this, all of our regressions include an indicator variable for each radial-bin.

Third, our three instruments are an indicator for outside, $\mathbb{1}(\text{Outside})$, this indicator interacted with the distance to the boundary, $\mathbb{1}(\text{Outside})x$, and this indicator interacted with meters of climbing required to reach the basin divide, $\mathbb{1}(\text{Outside})\Delta h$. We are concerned that horizontal displacement has a direct effect on sewer share. Given this, we control for horizontal displacement in three different ways. First, by including horizontal displacement as a control. Second, by including horizontal displacement interacted with segment-bin indicators,

Table 2: Sewers and log tract population density

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. OLS									
Sewer share	0.8576***	0.8576***	0.8576***	0.8019***	0.8019***	0.8019***	0.7397***	0.7397***	0.7397***
	(0.0259)	(0.0259)	(0.0259)	(0.0265)	(0.0265)	(0.0265)	(0.0264)	(0.0264)	(0.0264)
2. First-stage									
Outside	-0.0078**	-0.0055	-0.0051	-0.0143***	-0.0037	-0.0049	-0.0098***	-0.0023	-0.0046
	(0.0036)	(0.0040)	(0.0040)	(0.0037)	(0.0044)	(0.0044)	(0.0037)	(0.0047)	(0.0047)
x*Outside	-0.0001***		-0.0001***	-0.0001***		-0.0001***	-0.0001***		-0.0001***
	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)
ΔElev*Outside		-0.0002**	-0.0001		-0.0004***	-0.0004***		-0.0003**	-0.0002*
		(0.0001)	(0.0001)		(0.0001)	(0.0001)		(0.0001)	(0.0001)
3. IV *log(pop density)*									
Sewer share	1.8205***	2.9299**	2.0196***	3.8170***	1.9875**	3.9573***	6.0387***	0.4847	5.9413***
	(0.3540)	(1.4079)	(0.3568)	(0.4469)	(0.8654)	(0.4331)	(0.6026)	(1.5324)	(0.5846)
4. SATE *log(pop density)*									
Sewered	2.3389***	4.8942***	2.6074***	3.7429***	2.9344***	4.0004***	5.7074***	4.0916**	5.7297***
	(0.3641)	(1.2770)	(0.3662)	(0.3681)	(0.8500)	(0.3562)	(0.3850)	(1.4261)	(0.3774)
N	50039	50039	50039	50039	50039	50039	50039	50039	50039
F	93.22	6.617	63.10	84.28	19.78	62.25	73.25	5.924	50.32
Elevation	Y	Y	Y	Y	Y	Y	Y	Y	Y
π-bins	Y	Y	Y	Y	Y	Y	Y	Y	Y
x	Y	Y	Y						
seg×x				Y	Y	Y			
π-bins×x							Y	Y	Y

Note: *Sample is described by column 2 of table 1. Robust standard errors in parentheses.*
** $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

and finally, by including the interaction of the radial-bin indicator with the horizontal displacement, that is, $\mathbb{1}(\pi\text{-bin})_{jk}x_{jk}$. Because we also include a radial-bin indicator, in this final specification all of our parameter estimates are conditional on a radial-bin specific slope and intercept.

Finally, we are also concerned that elevation has an independent effect on sewer share and outcome, and so we also control for elevation. If vertical rise to the basin divide is more costly outside the basin than inside, and recalling that $\Delta h_{jk} \geq 0$, we expect α_2 to be negative. That is, the decline in sewer share should be more rapid as we move right from the basin divide in figure 8 than when we move left.

Table 2 presents our main set of estimation results. The top panel gives results of OLS regressions of equation (2) with different controls. Panel 2 presents first-stage regressions, equation (1), using different combinations of instruments and controls. Panel 3 presents TSLS estimates of equation (2) using the instruments and controls common to other results in the same column. We postpone discussion of panel 4. The bottom panel of the table describes the controls used in each specification, gives the sample size, and an F-statistic for

the instruments in the first-stage regression.

The columns are in three groups of three. The first three columns (1-3) present results for specifications that control for tract elevation, tract horizontal distance to the basin divide, and radial-bin intercepts. The second group of three (4-6) allows the effect of horizontal distance to vary by segment-bin, and the third group of three (7-9) allows the effect of horizontal distance to vary by radial-bin. The third group of three has the most flexible controls for potential confounding trends in horizontal distance and is our preferred specification.

Within each group of three columns, we vary the instruments that we use. In column 1 of each set (1,4,7), our instruments are the outside indicator and the outside indicator interacted with horizontal distance. In these regressions, causal identification relies on changes in elevation around the central basin boundary that results from horizontal displacement.

In column 2 of each set (2,5,8) our instruments are the outside indicator and the interaction of vertical distance to the basin divide and the outside indicator. In these regressions, identification relies on changes in elevation around the central basin boundary holding horizontal distance constant. The coefficient on these variables tells us the amount by which sewer share or log population density decrease with each additional meter of climbing to reach the basin divide for tracts outside the central basin. The third column of each set of three (3,6,9) includes all three instruments.

In the top panel of table 2 we see that a 1 percentage point increase in sewer share is associated with about a 1% increase in population density. This effect is estimated precisely and is stable across specifications. This about matches the pattern that figure 6 shows in the raw data. The OLS specification is identical for each of the three columns within a group, e.g., columns (1-3), because it is not affected by changes to the instrument set.

Panel 2 presents first-stage results. Point estimates for instruments have the expected negative signs. This confirms our intuition about the cost structure for providing sewer service. Whether measured by horizontal or vertical displacement, being on the outside of central basin divide increases the cost of sewer access and decreases its prevalence. The coefficients on the two interaction variables are stable across specifications and are estimated precisely. The relevant F-statistic is above the threshold for conventional weak instrument tests,

except in columns 2 and 8. These two columns both use the outside indicator and the outside indicator interacted with vertical displacement to the divide as instruments. Given this, we discount estimates based on these first-stage estimates, although we note that the weak instrument tests only apply to models with homogeneous treatment effects.

Finally, panel 3 presents our TSLS results. Ignoring columns 2 and 8, estimates range between about 1.8 in column 1 and 6.0 in column 7. Treatment effect estimates generally increase as we add controls. Column 9 has the most exhaustive set of controls and is our preferred estimate at 5.94. Comparing the TSLS elasticities of panel 3 with the OLS elasticities in panel 1, this suggests that the causal effect of sewer access is between about two and six times as large as the correlation in the raw data.

Positing that our instruments are as good as random conditional on controls, variation in the TSLS estimates is consistent with a heterogeneous treatment effects model where different controls and instruments lead to different weighted averages of treatment effects.

8 Instrument validity

The top panel of figure A1 provides evidence for the validity of our research design. This panel is a binscatter plot showing mean distance to the CBD net of radial-bin means. As expected, this plot is continuous and smooth. This is intuitive, but not guaranteed. If we had seen a step or kink in this distance gradient when we crossed the basin divide, it would have indicated a problem with our sampling rule.

The bottom two panels of figure A1 describe placebo tests. There are many places where sewer share does not vary across the basin divide. These places fall into two categories. Those that are sufficiently undesirable that they are unsewered on both sides of the basin divide, and those that are sufficiently desirable that they are completely sewered on both sides of the divide. If, in violation of the exclusion restriction, crossing a central basin divide has an independent effect on population density, then we should see this effect in these samples of tracts. The bottom two panels of figure A1 plot log tract population density as a function horizontal displacement to the basin divide. In panel (b) we restrict attention to the subset of radial-bins where tract mean sewer share is

above 90% within 2KM inside and outside of the basin divide. Panel (c) is similar, but restricts attention to radial-bins where tract mean sewer share is below 10% within 2KM inside and outside of the basin divide. That is, panels (b) and (c) plot how population density changes when we cross basin divides where sewer share does not vary. These figures are noisy, but do not support the hypothesis that basin divides affect population density independent of sewer share.

It is possible that the basin boundary is an important geological feature and that crossing it affects sewer share and outcomes because it impedes the movement of people and goods along with wastewater. The top panel of figure 7 (and the bottom panel of figure 5) suggests that this is rarely the case. Traveling 2KM inside from an average basin divide involves a drop of only about 30m, an average grade of about 1:70. This is suggestive, but could conceal dramatic local variation. The whiskers in both figures describe 95% CIs of each distance bin mean. That these confidence intervals are so tight suggests that local elevation profiles seldom differ much from the average. Thus, the raw data suggest that basin divides dramatic enough to impede the movement of goods and people are rare.

To investigate the possibility that our results are driven by rare bins where the basin divide is a dramatic geological feature, Appendix Table A2 replicates table 3 excluding radial-bins that contain a tract whose centroid is more than 100m below the basin divide. Relative to table 3, point estimates of TSLS treatment effect and SATE fall slightly in all specifications. However, in no specification does the change of sample result in a large enough change to coefficients that we fail to reject the hypothesis of no difference at standard confidence levels.

It is also possible that jurisdictional boundaries follow basin divides. If so, then crossing basin divides could affect sewer provision and population density because it involves crossing from one administrative unit to another. In our census data we can impute each tract's municipality from on its census identification code, and then check if municipal boundaries tend to follow basin divides.[5] To test the extent that municipal jurisdictions follow basin boundaries,

[5]For example, in Colombia the first 5 digits of the identification code for a *manzana* or census block is that manzana's municipal identification code. Similarly, for Brazil, Jordan, South Africa, and Tanzania, we retrieve a tract's administrative unit from the census information provided for each country.

we identify radial-bins where any of the tracts within 250 meters of the basin divide are in different municipalities. In our estimation sample only 3.5% of the radial-bins (containing 5.6% of tracts) contain municipal boundaries close to the basin divide. Basin divides rarely coincide with municipal boundaries.

Nevertheless, to investigate the possibility that our results are driven by the rare radial-bins where municipal boundaries and basin divides are close to each other, in Appendix Table A3 we replicate table 3 excluding the 3.5% of radial-bins where any of the tracts within 250 meters of the basin divide are in different municipalities. Point estimates of TSLS treatment effects fall slightly in all specifications. However, in no specification does the change of sample result in a large enough change to coefficients that we fail to reject the hypothesis of no difference at 5% confidence.

9 Estimating treatment effects from aggregate data

The econometric model described by equations (1) and (2) is an instrumental variables estimation with parametric controls and a continuous treatment. This poses two challenges to the causal interpretation of the treatment coefficient under effect heterogeneity.

First, Słoczyński (2021) and Blandhol et al. (2022) both investigate the properties of the TSLS estimator with linear additive covariates and binary treatment. Both conclude that this model leads to a weighted average of treatment effects that allows a causal interpretation only under restrictive conditions. Indeed, Blandhol et al. (2022) argues that these conditions are so strict as to be impossible to satisfy in practice.

Second, our problem involves a continuous (not binary) treatment. Whereas the binary case requires that we consider only two potential outcomes, treated and not, with a continuous treatment, we must consider a continuum of counterfactual outcomes for each unit. Chesher (2003) and Imbens and Newey (2009) consider the problem of estimating causal effects in a model with a continuous treatment, although their approaches are challenging when multi-dimensional covariates are present.

Angrist et al. (2000) and Kolesár and Plagborg-Møller (2024) study causal interpretation of linear TSLS with a continuous treatment. They show that the TSLS estimand can be interpreted as a weighted average of the marginal effects,

although this average is difficult to interpret and not of obvious economic interest. Furthermore, the weights can be negative for some specifications and assumptions about the first-stage equation.

This section develops a framework for estimating the MTE model with parcel-level binary treatment using data aggregated to the tract level. We build on the MTE model of Carneiro et al. (2011) which permits the estimation of treatment effects in instrumental variables estimations with parametric controls and binary treatment. We start with the observation that our treatment is binary at the parcel level, so that if we can estimate an MTE model at the parcel level, we can allow for parametric controls and parcel level heterogeneity in treatment effects and the propensity to select sewage access.

The obvious obstacle is that our data describes census tract aggregates, not the underlying parcel level observations themselves. We resolve this problem by describing a small variance approximation to the parcel level MTE model, and then taking expectations of this approximation over tracts. This leads to tract level estimating equations that identify the causal effects defined in the parcel-level MTE model. This logic requires that we take tract level averages of non-linear parcel level equations and it leads to tract level estimating equations that depend on the within-tract variance of the parcel level characteristics. While we do not observe any parcel level data, our data reports within tract variances of the controls, and so our data allows us to estimate this model to recover tract level means of parcel level treatment effects.

We begin by stating the parcel level model. Let i index parcels and j index census tracts. Y_{ij} is the outcome variable of interest for parcel i in tract j, D_{ij} is our treatment variable and takes the value one if a parcel has sewer access, and zero if not. X_{ij} is a vector of covariates, and Z_{ij} a vector of instruments. Let $W_{ij} = (X_{ij}, Z_{ij})$. Controls and instruments are the same as in the earlier TSLS estimates of equations (1) and (2).

Our parcel level MTE model consists of two linear equations describing the relationship between controls and potential outcomes,

$$Y_{ij}(1) = X_{ij}'\beta_1 + U_{ij}(1),$$
$$Y_{ij}(0) = X_{ij}'\beta_0 + U_{ij}(0),$$

along with a parcel level selection equation,

$$D_{ij} = 1\{p(W_{ij}) \geq V_{ij}\}.$$

As is standard in the MTE literature, the selection equation assumes additive separability of the unobserved heterogeneity, V_{ij}, and the latent utility term involving W_{ij}. Moreover, and without loss of generality, we normalize the unobserved parcel level heterogeneity in the selection equation, V_{ij}, to be uniform on the unit interval. As a result, $p(W_{ij})$ coincides with the parcel level propensity to select into treatment $\Pr(D_{ij} = 1|W_{ij})$.

We also impose the practical exogeneity condition of Carneiro et al. (2011), $(U_{ij}(1), U_{ij}(0), V_{ij}) \perp W_{ij}$. This implies that unobservable heterogeneity in the parcel level selection and potential outcome equations is statistically independent of parcel level observables. In this case, the local IV regression is given by

$$Y_{ij} = X'_{ij}\beta_0 + p(W_{ij})X'_{ij}(\beta_1 - \beta_0) + \phi(p(W_{ij})) + U_{ij}, \tag{3}$$

where $\phi(\cdot)$ is a control function describing the dependence between $(U_{ij}(1), U_{ij}(0))$ and V_{ij}. The parcel level marginal treatment effect (MTE) conditional on $X_{ij} = x$ and $V_{ij} = v$ is the derivative of equation (3) with respect to the propensity score. That is,

$$\text{MTE}(x,v) \equiv E[Y_{ij}(1) - Y_{ij}(0)|X_{ij} = x, V_{ij} = v] = x'(\beta_1 - \beta_0) + \phi'(v). \tag{4}$$

Because V_{ij} is uniform, the population ATE is,

$$\text{ATE} = E[\text{MTE}(X_{ij},V_{ij})] = E[X_{ij}]'(\beta_1 - \beta_0) + \int_0^1 \phi'(v)dv. \tag{5}$$

Under the practical exogeneity condition, we have X_{ij} independent of V_{ij} so that CATE (Conditional Average Treatment Effect) given X_{ij} is linear in X_{ij},

$$\text{CATE}(X_{ij}) = E[\text{MTE}(X_{ij},V_{ij})|X_{ij}] = X'_{ij}(\beta_1 - \beta_0) + \int_0^1 \phi'(v)dv. \tag{6}$$

$\text{CATE}(X_{ij})$ gives the treatment effect for an average unit, here a parcel, with observable characteristics X_{ij}.

Let $\sigma^2 R_j$ denote the variance-covariance matrix of W_{ij} within each tract j, with σ^2 a scaling term that we introduce to facilitate the small-variance

approximation we describe shortly. Our data reports tract averages, $(\overline{Y}_j, \overline{D}_j, \overline{W}_j)$, and $\sigma^2 R_j$. We discuss the calculation of $\sigma^2 R_j$ below.

To estimate a parcel level model from tract means, $(\overline{Y}_j, \overline{D}_j, \overline{W}_j)$, and the within tract variances of W_{ij}, $\sigma^2 R_j$, we make three additional assumptions. First, that the parcel level propensity score $p(W_{ij}) = \Pr(D_{ij} = 1|W_{ij})$ is three times continuously differentiable. Second, that the third moments of W_{ij} exist. Third, that $\phi(p)$ is quadratic,

$$\phi(p) = \alpha_0 + \alpha_1 p + \frac{1}{2}\alpha_2 p^2. \tag{7}$$

Given these assumptions, Appendix C derives a tract-level regression equation that estimates the parcel level MTE model.

If we further assume a linear probability model for the propensity score,

$$p(W_{ij}) = W_{ij}'\gamma. \tag{8}$$

then we can estimate the parameters $(\beta_1 - \beta_0)$, α_0, α_1, and α_2 from the parcel level first-stage and structural equations (3) and (8), with the following two tract level estimating equations,

$$p(\overline{W}_j) = \overline{W}_j'\gamma, \tag{9}$$

$$\begin{aligned} \overline{Y}_j = {} & \overline{X}_j'\beta_0 + p(\overline{W}_j)\overline{X}_j(\beta_1 - \beta_0) + \sigma^2(\beta_1 - \beta_0)k_{1j} \\ & + \alpha_0 + \alpha_1 \cdot p(\overline{W}_j) \\ & + \frac{1}{2}\alpha_2 \cdot [p(\overline{W}_j)^2 + \sigma^2 k_{3j}] + O_j(\sigma^3) + \eta_j, \end{aligned} \tag{10}$$

where k_{1j} and k_{3j} are observable scalars calculated from $\overline{W}_j$, R_j and $\widehat{\gamma}$, and $O_j(\sigma^3)$ is a tract level approximation error that vanishes as $\sigma^3 \to 0$.

Written this way, we see that these are just the original MTE structural equation estimated on tract averages, but with the addition of terms involving the variances of tract level variables. We also see that, even though some of the regressors involve non-linear calculations, equation (10) is linear in parameters. Hence, assuming that the approximation error term $O_j(\sigma^3)$ is negligible, we can estimate equations (9) and (10) with OLS to identify $(\beta_1 - \beta_0)$, α_0, α_1, and α_2. This lets us estimate MTE(x,p) and other causal estimands by plugging these parameters into (4) and (6).

We obtain the tract level sample average treatment effects by averaging $\text{CATE}(X_{ij})$ over all parcels in each tract j. Under the practical exogeneity assumption this leads to the following expression for the sample average treatment effect for a tract with mean parcel observables $\overline{X}_j$,

$$\begin{aligned}\text{SATE}_j &= \overline{X}_j'(\beta_1 - \beta_0) + \int_0^1 \phi'(v)dv \\ &= \overline{X}_j'(\beta_1 - \beta_0) + \alpha_1 + \frac{1}{2}\alpha_2. \end{aligned} \tag{11}$$

We discuss the calculation of standard errors in Appendix C.

Recall that our identification strategy relies heavily on dummy variables at the city, segment, or radial-bin level, along with their interaction with the horizontal distance to the basin divide. This creates two practical problems. First, with so many regressors, our estimators are difficult to compute, particularly if we rely on a non-linear, e.g., Logit, functional form for the propensity score $p(\cdot)$. Second, the incidental parameter problem arises in the estimation of fixed effects in nonlinear regressions; see e.g., (Lancaster, 2000). These two problems motivate our reliance on the linear probability model, equation (9).

That we rely on a large set of controls creates another problem. The statement of the MTE model allows for treatment effects to vary with unobserved resistance to treatment and with observed characteristics, the term $p(W_{ij})X_{ij}'(\beta_1 - \beta_0)$ in (3). Here, the number of coefficients also increases linearly in the dimension of X_{ij}. Given the large number of fixed effects in our regression equations, this specification is an extremely flexible description of treatment heterogeneity. As a practical matter, estimates of the $(\beta_1 - \beta_0)$ coefficients are difficult to compute and unstable, and so we impose the restriction $(\beta_1 = \beta_0)$ in all of the results that we present. This restricts attention to the case where treatment effects are heterogeneous on unobservables only.

To estimate this model, we require data describing the within tract variances of parcel level control variables and instruments. The list of instruments and controls that we use in our estimates of equation (2) involves just four variables and interactions of these variables; elevation, horizontal distance, an outside indicator, and radial-bin indicators. By construction, radial-bin indicators are constant within a tract, and so their within tract variance is trivially zero. This leaves elevation, horizontal distance, and the outside indicator.

Our census data does not report any data at the parcel level. However, our elevation data is gridded data with a spatial resolution of about 30 meters, not much larger than a parcel. We can also evaluate the outside indicator and horizontal distance for each grid cell in the elevation data. Finally, for all of our census data, we have tract boundary files. Putting these data together, we can calculate the within tract variances and covariances of all of our control variables and instruments for the universe of tracts in our study area. To implement the estimator described by equations (9) and (10), we use within tract calculations of variances based on 30m grid cells to proxy for parcel values, and estimate both equations with OLS.

With these estimates in hand, we can then evaluate the SATE_j described in equation (11), and then average over the whole sample of tracts to get an estimate for the average treatment effect. The resulting estimates are reported in panel 4 of table 2, along with standard errors calculated as we describe in Appendix C.

Ignoring columns 2 and 8 where the instruments are weak, the range of SATE estimates is from about 2.3 to 5.7, narrower than the about 1.8 to 6.0 range of estimates for TSLS. The SATE estimates and the TSLS estimates are close compared to the precision of the estimates, particularly for our preferred specification in column 9, and the SATE estimates are marginally more precise than the TSLS estimates. Unlike the TSLS estimations, the SATE estimate in column 8 is not an outlier.

While it is tempting to interpret the TSLS estimations as LATES, this interpretation rests on strong assumptions, and the precise formulation of the resulting regression weighted LATE is not obviously of economic interest. Our SATE, on the other hand, has a straightforward interpretation. It describes the amount by which population density changes on an average parcel when that parcel receives sewer access.

10 Sewers and sorting

We now investigate whether sewers also affect the demographic characteristics of residents. This is of intrinsic interest and helps us to understand the incidence of the benefits of sewer access. Do sewers help incumbent residents, or do they precipitate the arrival of more affluent migrants?

Table 3: Sewers and log tract income

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. OLS									
Sewer share	0.4701***	0.4701***	0.4701***	0.4141***	0.4141***	0.4141***	0.4007***	0.4007***	0.4007***
	(0.0140)	(0.0140)	(0.0140)	(0.0139)	(0.0139)	(0.0139)	(0.0144)	(0.0144)	(0.0144)
2. First-stage									
Outside	-0.0100*	-0.0193***	-0.0191***	-0.0170***	-0.0155***	-0.0165***	-0.0150***	-0.0148**	-0.0161**
	(0.0052)	(0.0054)	(0.0054)	(0.0053)	(0.0060)	(0.0060)	(0.0054)	(0.0063)	(0.0063)
x*Outside	-0.0000***		-0.0000***	-0.0000***		-0.0000***	-0.0000***		-0.0000***
	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)
ΔElev*Outside		0.0005***	0.0005***		-0.0001	-0.0000		0.0000	0.0001
		(0.0001)	(0.0001)		(0.0001)	(0.0001)		(0.0001)	(0.0001)
3. IV *log(income)*									
Sewer share	2.8644***	1.7817***	2.6017***	1.5259***	-1.0608	1.5252***	0.8452**	-1.9792*	0.8242**
	(0.8914)	(0.4943)	(0.4808)	(0.4589)	(0.7833)	(0.4591)	(0.3678)	(1.1516)	(0.3649)
4. SATE *log(income)*									
Sewered	2.9623***	1.7721***	2.6010***	1.5356***	-1.0600	1.5340***	0.6886*	-1.7475*	0.6654
	(0.4819)	(0.4382)	(0.3427)	(0.3589)	(0.6594)	(0.3602)	(0.3501)	(0.7790)	(0.3482)
N	25424	25424	25424	25424	25424	25424	25424	25424	25424
F	93.22	6.617	63.10	84.28	19.78	62.25	73.25	5.924	50.32
Elevation	Y	Y	Y	Y	Y	Y	Y	Y	Y
π-bins	Y	Y	Y	Y	Y	Y	Y	Y	Y
x	Y	Y	Y						
seg×x				Y	Y	Y			
π-bins×x							Y	Y	Y

Note: *Sample is described by column 2 of table 1, but includes only the two countries, Brazil and South Africa, that report income in their census. Robust standard errors in parentheses.*
** $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

We investigate this question in two ways. First, the Brazilian and South African censuses report tract income directly. Thus, for these two countries, we implement exactly the same research design that we used to understand the effect of sewers on population density. Table 3 reports results. The organization of table 3 is identical to table 2, except that the dependent variable in all cases is the log of tract mean income. The top panel reports OLS regressions of equation (2). Panel 2 reports first-stage estimates, equation (1). Panel 3 reports TSLS estimates of equation (2). Panel 4 reports the SATE described by equation (11).

Recalling that the dependent variable of interest is a share and the dependent variable is a logarithm, the OLS results indicate that a 1 percentage point increase in sewer share is associated with about 0.4% increase in income. Estimates in panels three and four are somewhat larger. In our preferred estimate in column 9, a 1 percentage point increase in sewer share causes about 0.8% increase in tract mean income.

From table 1, we see that the mean and standard deviation of income are both about 950 dollars. Abusing the marginal nature of our result, providing universal sewer access to a previously unsewered tract increases income by

Table 4: Sewers and tract literacy rate

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. OLS									
Sewer share	0.0022	0.0022	0.0022	-0.0044	-0.0044	-0.0044	0.0228***	0.0228***	0.0228***
	(0.0176)	(0.0176)	(0.0176)	(0.0183)	(0.0183)	(0.0183)	(0.0031)	(0.0031)	(0.0031)
2. First-stage									
Outside	-0.0079**	-0.0057	-0.0053	-0.0144***	-0.0038	-0.0050	-0.0099***	-0.0023	-0.0046
	(0.0036)	(0.0040)	(0.0040)	(0.0037)	(0.0044)	(0.0044)	(0.0037)	(0.0047)	(0.0047)
x*Outside	-0.0001***		-0.0001***	-0.0001***		-0.0001***	-0.0001***		-0.0001***
	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)
ΔElev*Outside		-0.0002**	-0.0001		-0.0004***	-0.0004***		-0.0003**	-0.0002*
		(0.0001)	(0.0001)		(0.0001)	(0.0001)		(0.0001)	(0.0001)
3. IV *literacy rate*									
Sewer share	0.1238**	-0.4413**	0.1125*	0.0405	0.0637	0.0616*	0.0288	0.0345	0.0377
	(0.0600)	(0.2066)	(0.0596)	(0.0396)	(0.0828)	(0.0338)	(0.0331)	(0.0867)	(0.0317)
4. SATE *literacy rate*									
Sewered	0.1669**	-0.2995*	0.1620*	0.0442	0.0246	0.0596	0.0339	-0.0036	0.0408
	(0.0628)	(0.1470)	(0.0633)	(0.0375)	(0.0698)	(0.0330)	(0.0315)	(0.0911)	(0.0304)
N	50019	50019	50019	50019	50019	50019	50019	50019	50019
F	93.22	6.617	63.10	84.28	19.78	62.25	73.25	5.924	50.32
Elevation	Y	Y	Y	Y	Y	Y	Y	Y	Y
π-bins	Y	Y	Y	Y	Y	Y	Y	Y	Y
x	Y	Y	Y						
seg×x				Y	Y	Y			
π-bins×x							Y	Y	Y

Note: *Sample is described by column 2 of table 1, but excludes Jordan, for which no data on literacy or education is available. Robust standard errors in parentheses.*
** $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

about 82%. This is a bit less than one standard deviation. This suggests that sewer access has only a modest effect on the spatial distribution of income groups across neighborhoods.

We next investigate whether sewers affect the literacy rate of tract residents. The censuses for Brazil, Colombia, South Africa and Tanzania report on educational attainment or literacy such that we can create a standardized measure of the share literate in each tract. We now ask how this measure varies with sewer access using the same approach as we use to investigate the effect of sewer access on population density. Table 4 reports these results. OLS results suggest that literacy rates are about 0.02 percentage points higher in a tract when the sewer share is 1 percentage point higher. IV results, in panels three and four, suggest a smaller effect. In our preferred specification of column 9, a 1 percentage point increase in sewer share causes about a 0.04 percentage point increase in the literacy rate, although this effect is not distinguishable from zero.

Taking the 0.04 percentage point IV estimate of column 9 seriously, providing universal access to a completely unsewered tract increases the literacy rate by about 4 percentage points. From table 1 the tract mean literacy rate is about 92%

with a standard deviation of about 40%. This suggests that sewer access plays at most a small role in spatial distribution of literacy.

Our estimates of the effects of sewer access on income and literacy do not suggest that improved sewer access in a tract precipitates the arrival of more affluent migrants nor the displacement of current residents. On the contrary, our results suggest that improving sewer access in tract will provide sewer service to incumbents or to demographically similar immigrants.

11 Discussion

Sewers and urban density

Results so far establish an effect size. Using our preferred estimate from column 9 of table 2, adding 1% of sewer connections to a tract causes and increase in tract population density of about 6%. It is not immediately clear whether this is an economically important effect. We would like to develop some intuition around this issue.

Define a "city" as consisting of all census tracts inside or within 2KM of the central basin. This is the sample of cities and tracts described in column 1 of table 1. For each city in our sample, consider a counterfactual case where we add 1% to the count of sewer connections in the city. We add these connections, tract by tract, by first sewering all unsewered households in the most densely populated tract where sewer access is not universal. If completing sewer coverage in this tract does not exhaust the 1% increase in total connections, we move on to the next most densely populated tract containing unsewered households, and so on, until we allocate all of the 1% of new connections.

For each city, this process results in a counterfactual city where a subset of tracts has better sewer access than in the observed case. We can then use our estimates of treatment effects to calculate the implied increase in population in these tracts. Assume that the 1 percentage point increase in sewer connections increases city population by inducing rural residents to migrate to the city, and a 6% treatment effect. In this case, mechanically, our counterfactual cities house 6% more people than their observed counterparts.

This effect seems large in the following sense. Baum-Snow (2007) finds that each radial interstate highway decreased the density of US central cities by 9%. Our estimates suggest the opposite effect can be accomplished by adding about

1.5% to a central city's stock of sewer connections. That is, a 1% increment to a central city's sewer share is about two thirds as important for urban form as is a single limited access radial highway.

The increase in person weighted density is also of interest, but must be evaluated tract by tract. We perform this calculation for all cities in our estimation sample. Figure 9 presents a histogram summarizing our results. The modal city in our sample experiences an about 17% increase person weighted density, and the upper tail experiences much larger increases.

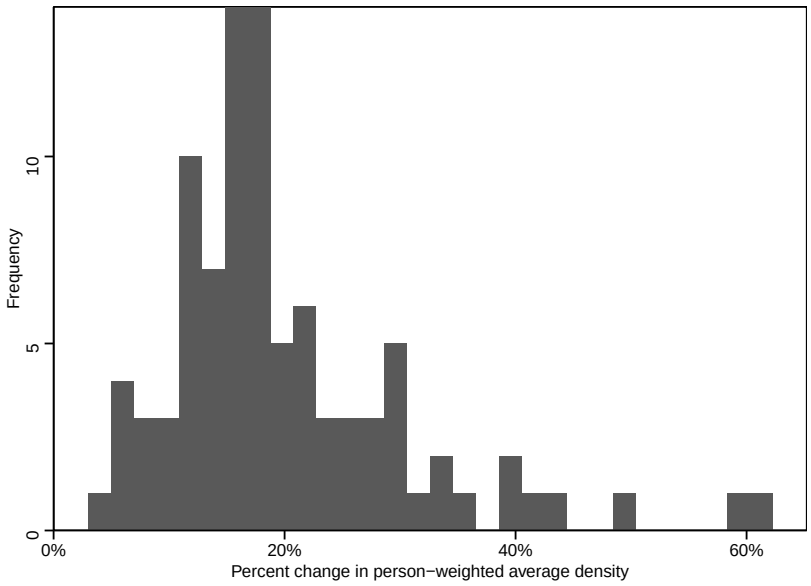
These effects also seem large. The relationship between density and labor productivity is well established, and a central estimate is that doubling the density of a city increases wages by about 5%. Combining this estimate with the modal 17% increase in person weighted density that we see in figure 9, suggests that adding 1% to the stock of sewer connections will increase wages of incumbent residents by about 0.85%. This is a flow. Taking its discounted present value using a 5% interest rate gives about 17% of the city's total annual wage bill. We suspect that this benefit alone will often be large relative to the cost of adding the required 1% of connections. Including the likely health benefits to incumbents and the likely wage increase experienced by rural migrants will increase this estimate of benefits further.

CBD access and sewer access

It is now common to evaluate public transit systems at least partly on the basis of the extent to which they improve access to the central city and thereby improve the functioning of the labor market, e.g., Tsivanidis (2019), Zárate (2022), Heblich et al. (2020). By facilitating higher residential densities, improving the sewer network within walking distance of the CBD also improves access to the CBD.

To assess the importance of this effect, we ask how many people would gain access to the CBD if we completed the sewer network in all tracts with centroids within 4KM of the CBD, and then allowed density to adjust, keeping the total population of the city constant. Using our same 6% treatment effect, we find that completely sewering the area within 4KM of the CBD increases the share of the city's population within walking distance of the center by 18% for the average city in our sample.

Figure 9: Histogram of counterfactual changes in density



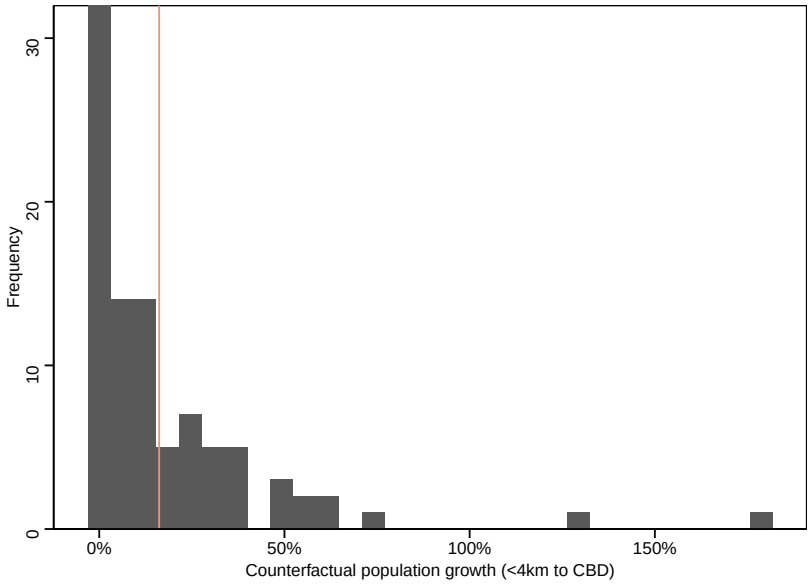
Note: *Histogram of mean increase in person weighted density by city that results from adding 1% more sewer connections to the most densely populated tracts where access is not universal.*

This 18% mean value conceals cross-city heterogeneity. Figure 10 is a histogram of describing the magnitude of the effect. From Tsivanidis (2019), the Transmilenio BRT allows about 18% of the population of Bogota to access the CBD. Figure 10 suggests that building out the central city sewer network has an effect equal to or larger than that of a successful BRT system in 32% of our sample cities.

12 Conclusion

We estimate the effects of sewer access on population density in a sample of developing world cities using a novel identification strategy that derives from principles of wastewater engineering. We estimate that a 1 percentage point increase in the share of tract households with sewer access causes about a 6% change in tract population and only a small effect on the tract mean income and literate share. This suggests that sewer construction projects impact people like the incumbent residents and do not precipitate the arrival of more affluent migrants.

Figure 10: Percentage of city population gaining walking access to the CBD when central city sewer network in completed.



Note: *Histogram of showing the change in the percentage of city population that would gain walking access to the* CBD *if all tracts within 4*KM *of the* CBD *had sewer access increased to 100% and the consequent population increase was from more remote tracts. Vertical red line at 18% is the share of Bogota's population that gained access to the* CBD *because of the the Transmilenio* BRT *system.*

Because the treatment we consider, share of tract households with sewer access, is continuous, the interpretation conventional TSLS estimates is difficult except under strong assumptions. To arrive at an estimand that can be interpreted as an average treatment effect in a heterogeneous treatment effects framework, we note that at the parcel level, our treatment is binary. This means that a MTE/LIV estimation can yield estimates of parcel level average treatment effects. We develop a technique to estimate this parcel level model from tract level data using a small variance approximation. Thus, in addition to the difficult to interpret TSLS estimate of treatment effects, we also estimate a Sample Average Treatment Effect of sewer access on parcel level outcomes. In practice, both approaches lead to similar estimates of effect size. In addition to our contribution to understanding the effects of sewers on urbanization, we hope that this technique will prove useful to other researchers faced with similar estimation problems.

Finally, we perform two simple counterfactual experiments to assess whether our estimated 6% effect size is economically important. The first of these suggests that adding 1% to a city's sewer connections in its densest neighborhoods has an effect on population density about 2/3 as large and of opposite sign as as a single radial interstate highway ray. The second experiment suggests that completely building out the sewer network within 4KM of the CBD will increase the share of city population living in this disk by 18% on average, and by much more in some cities. This means that building out central city sewer access is often as important for improving access to central city labor markets as is a successful BRT system.

References

Aldous, D. (1999). *International turf management handbook*. CRC Press.

Allcott, H., Collard-Wexler, A., and O'Connell, S. D. (2016). How do electricity shortages affect industry? evidence from india. *American Economic Review*, 106(3):587–624.

Alsan, M. and Goldin, C. (2019). Watersheds in child mortality: The role of effective water and sewerage infrastructure, 1880–1920. *Journal of Political Economy*, 127(2):586–638.

Anderson, D. M., Charles, K. K., and Rees, D. I. (2018). Public health efforts and the decline in urban mortality. Technical report, National Bureau of Economic Research.

Angrist, J., Graddy, K., and Imbens, G. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies*, 67(3):499–527.

Ashraf, N., Glaeser, E., Holland, A., and Steinberg, B. M. (2017). Water, health and wealth. Technical report, National Bureau of Economic Research.

Baum-Snow, N. (2007). Did highways cause suburbanization? *The Quarterly Journal of Economics*, 122(2):775–805.

Bhalotra, S. R., Diaz-Cayeros, A., Miller, G., Miranda, A., and Venkataramani, A. S. (2021). Urban water disinfection and mortality decline in lower-income countries. *American Economic Journal: Economic Policy*, 13(4):490–520.

Blandhol, C., Bonney, J., Mogstad, M., and Torgovitsky, A. (2022). When is TSLS actually LATE? Technical report, National Bureau of Economic Research.

Brazilian Institute of Geography and Statistics (2012). Brazil demographic census 2010. Technical report. Rio de Janeiro, Brazil, 2012.

Carneiro, P., Heckman, J. J., and Vytlacil, E. (2011). Estimating marginal returns to education. *American Economic Review*, 101(6):2754–2781.

Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78(3):451–462.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, 71(5):1405–1441.

Coury, M., Kitagawa, T., Shertzer, A., and Turner, M. (2022). The value of piped water and sewers: Evidence from 19th century Chicago. Technical report, National Bureau of Economic Research.

Department of Statistics (Jordan) (2015). Jordan population and housing census 2015. Technical report. Department of Statistics, Aman Jordan.

Ferrie, J. P. and Troesken, W. (2008). Water and Chicago's mortality transition, 1850–1925. *Explorations in Economic History*, 45(1):1–16.

Galiani, S., Gertler, P., and Schargrodsky, E. (2005). Water for life: The impact of the privatization of water services on child mortality. *Journal of Political Economy*, 113(1):83–120.

Gamper-Rabindran, S., Khan, S., and Timmins, C. (2010). The impact of piped water provision on infant mortality in Brazil: A quantile panel data approach. *Journal of Development Economics*, 92(2):188–200.

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

Ghani, E., Goswami, A. G., and Kerr, W. R. (2016). Highway to success: The impact of the golden quadrilateral project for the location and performance of Indian manufacturing. *The Economic Journal*, 126(591):317–357.

Gibson, J., McKenzie, D., and Rohorua, H. (2014). Development impacts of seasonal and temporary migration: A review of evidence from the Pacific and Southeast Asia. *Asia & the Pacific Policy Studies*, 1(1):18–32.

Heblich, S., Redding, S. J., and Sturm, D. M. (2020). The making of the modern metropolis: evidence from London. *The Quarterly Journal of Economics*, 135(4):2059–2133.

Henderson, J. V. and Turner, M. A. (2020). Urbanization in the developing world: too early or too slow? *Journal of Economic Perspectives*, 34(3):150–73.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Jedwab, R. and Storeygard, A. (2022). The average and heterogeneous effects of transportation investments: Evidence from sub-saharan africa 1960–2010. *Journal of the European Economic Association*, 20(1):1–38.

Kesztenbaum, L. and Rosenthal, J. L. (2017). Sewers' diffusion and the decline of mortality: The case of Paris, 1880–1914. *Journal of Urban Economics*, 98:174–186.

Kolesár, M. and Plagborg-Møller, M. (2024). Dynamic causal effects in a nonlinear world: the good, the bad, and the ugly. *arXiv preprint arXiv:2411.10415*.

Lagakos, D., Marshall, S., Mobarak, A. M., Vernot, C., and Waugh, M. E. (2020). Migration costs and observational returns to migration in the developing world. *Journal of Monetary Economics*, 113:138–154.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95:391–413.

Lipscomb, M., Mobarak, A. M., and Barham, T. (2013). Development effects of electrification: Evidence from the topographic placement of hydropower plants in Brazil. *American Economic Journal: Applied Economics*, 5(2):200–231.

Mara, D. (1996). *Low-cost sewerage*. John Wiley London.

NASA JPL (2013). Nasa shuttle radar topography mission global 1 arc second [data set]. Technical report. Last accessed 19 May 2022.

NASA/METI/AIST/Japan Spacesystems and US/Japan ASTER Science Team (2019). ASTER global digital elevation model v003. Technical report. Last accessed 19 May 2022.

National Administrative Department of Statistics (2018). Colombia population and housing census 2018. Technical report. Bogotá, Colombia: National Administrative Department of Statistics.

National Bureau of Statistics (Tanzania), Office of the Chief Government Statistician (2012). Tanzania population and housing census 2012. Technical report. National Bureau of Statistics (Tanzania), Office of the Chief Government Statistician (Zanzibar).

Ogasawara, K. and Matsushita, Y. (2018). Public health and multiple-phase mortality decline: Evidence from industrializing Japan. *Economics & Human Biology*, 29:198–210.

Słoczyński, T. (2021). When should we (not) interpret linear IV estimands as LATE? *unpublished manuscript*.

Statistics South Africa (2011). Census 2011. Technical report. Pretoria, South Africa: Statistics South Africa.

UN DESA Population Division (2018). World urbanization prospects: the 2018 revision.

Tsivanidis, N. (2019). Evaluating the impact of urban transit infrastructure: Evidence from Bogota's Transmilenio.

Uuemaa, E., Ahi, S., Montibeller, B., Muru, M., and Kmoch, A. (2020). Vertical accuracy of freely available global digital elevation models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEM). *Remote Sensing*, 12(21):3482.

Zárate, R. D. (2022). Spatial misallocation, informality, and transit improvements: Evidence from Mexico City.

Appendix A Supplemental results

Table A1: Estimation sample by country

	(1) Cities	(2) π-bins	(3) Tracts	(4) Share inside	(5) Tract area	(6) Pop. Density	(7) Sewer share
Brazil	55	936	22,372	0.54	0.23	17,093	0.68
Colombia	17	285	23,059	0.53	0.07	29,049	0.72
Jordan	2	4	18	0.33	1.34	1,196	0.60
South Africa	12	211	3,053	0.52	0.31	8,635	0.78
Tanzania	6	77	1,537	0.66	0.07	25,803	0.08

Note: *Census data for Jordan reports households, not people. Columns (4-7) are tract weighted averages. Population density is people per* KM^2 *and tract area is* KM^2.

Table A2: Sewers and log tract population density, dropping hilly areas

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. OLS									
Sewer share	0.8000***	0.8000***	0.8000***	0.7604***	0.7604***	0.7604***	0.7063***	0.7063***	0.7063***
	(0.0258)	(0.0258)	(0.0258)	(0.0271)	(0.0271)	(0.0271)	(0.0269)	(0.0269)	(0.0269)
2. First-stage									
Outside	-0.0079**	0.0049	0.0041	-0.0112***	-0.0018	-0.0030	-0.0078**	-0.0047	-0.0066
	(0.0038)	(0.0047)	(0.0047)	(0.0039)	(0.0051)	(0.0051)	(0.0039)	(0.0054)	(0.0054)
x*Outside	-0.0001***		-0.0001***	-0.0001***		-0.0001***	-0.0001***		-0.0001***
	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)
ΔElev*Outside		-0.0007***	-0.0006***		-0.0004***	-0.0004**		-0.0001	-0.0001
		(0.0001)	(0.0001)		(0.0002)	(0.0002)		(0.0002)	(0.0002)
3. IV *log(pop density)*									
Sewer share	1.2437***	2.5149***	1.6045***	3.2809***	1.0840	3.4718***	5.2988***	-5.4821	5.3133***
	(0.3312)	(0.7396)	(0.3126)	(0.4867)	(1.0356)	(0.4775)	(0.6095)	(4.5492)	(0.6089)
4. SATE *log(pop density)*									
Sewered	1.6015***	3.4154***	2.0506***	3.1334***	1.8174	3.4180***	4.8707***	-0.6425	4.9112***
	(0.3468)	(0.6505)	(0.3314)	(0.4058)	(1.0024)	(0.3901)	(0.4093)	(2.6012)	(0.4084)
N	44399	44399	44399	44399	44399	44399	44399	44399	44399
F	90.69	18.37	67.77	60.59	8.946	42.90	61.82	1.472	41.26
Elevation	Y	Y	Y	Y	Y	Y	Y	Y	Y
π-bins	Y	Y	Y	Y	Y	Y	Y	Y	Y
x	Y	Y	Y						
seg$\times$x				Y	Y	Y			
π-bins$\times$x							Y	Y	Y

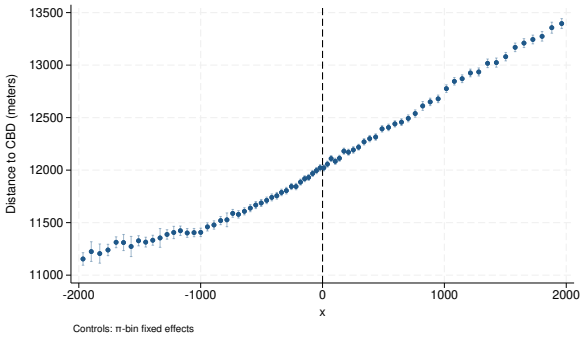
Note: *Sample is described by column 2 of table 1, but excludes radial-bins containing a tract centroid 100m or more below the basin divide. Robust standard errors in parentheses.*
** $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

Table A3: Sewers and log tract population density, dropping municipal borders

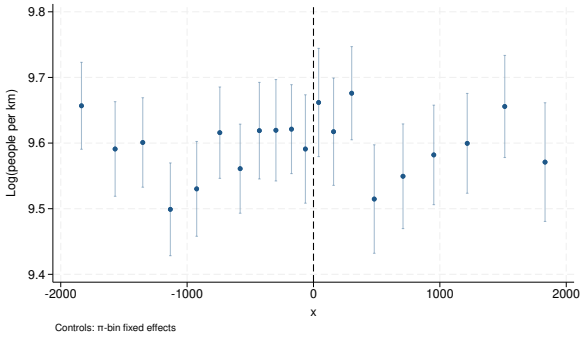
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. OLS									
Sewer share	0.8526***	0.8526***	0.8526***	0.7917***	0.7917***	0.7917***	0.7322***	0.7322***	0.7322***
	(0.0266)	(0.0266)	(0.0266)	(0.0276)	(0.0276)	(0.0276)	(0.0272)	(0.0272)	(0.0272)
2. First-stage									
Outside	-0.0110***	-0.0092**	-0.0087**	-0.0157***	-0.0087*	-0.0094**	-0.0157***	-0.0089*	-0.0108**
	(0.0037)	(0.0041)	(0.0041)	(0.0038)	(0.0045)	(0.0045)	(0.0038)	(0.0048)	(0.0048)
x*Outside	-0.0001***		-0.0001***	-0.0000***		-0.0000***	-0.0001***		-0.0001***
	(0.0000)		(0.0000)	(0.0000)		(0.0000)	(0.0000)		(0.0000)
ΔElev*Outside		-0.0002**	-0.0001		-0.0003***	-0.0003**		-0.0002**	-0.0002*
		(0.0001)	(0.0001)		(0.0001)	(0.0001)		(0.0001)	(0.0001)
3. IV *log(pop density)*									
Sewer share	1.3772***	0.8313	1.5311***	3.3969***	0.6798	3.7089***	4.9751***	-0.3898	4.9051***
	(0.3450)	(1.1370)	(0.3463)	(0.5949)	(1.0060)	(0.5886)	(0.5607)	(1.1567)	(0.5460)
4. SATE *log(pop density)*									
Sewered	1.8790***	2.4296*	2.0880***	3.3939***	2.0763*	3.8617***	4.6945***	2.6938*	4.7624***
	(0.3690)	(1.1650)	(0.3714)	(0.5129)	(1.0233)	(0.4990)	(0.3895)	(1.0862)	(0.3831)
N	47245	47245	47245	47245	47245	47245	47245	47245	47245
F	96.84	8.642	65.25	44.09	13.67	32.07	64.74	9.891	44.39
Elevation	Y	Y	Y	Y	Y	Y	Y	Y	Y
π-bins	Y	Y	Y	Y	Y	Y	Y	Y	Y
x	Y	Y	Y						
seg$\times$x				Y	Y	Y			
π-bins$\times$x							Y	Y	Y

Note: *Sample is described by column 2 of table 1, but excludes radial bins containing tract centroids closer to the basin divide than 250m but in different municipalities. Robust standard errors in parentheses.*
** $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

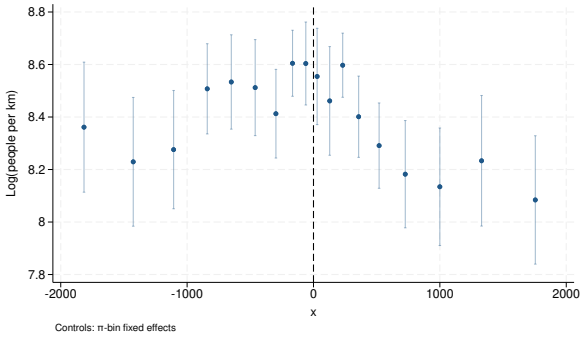
Figure A1: Placebo and balance tests



(a)



(b)



(c)

Note: *Each panel is a binscatter plot with horizontal displacement from the central basin divide on the x-axis. (a) Bin mean of distance to* CBD *net of radial-bin mean. That this plot is continuous reassures us that we have not introduced an unintended sampling restriction. (b) Bin mean population density for all radial-bins where tract mean sewer share is above 90% within* 2KM *of the basin divide. (c) Same as (b) but for radial-bins where tract mean sewer share is below 10% within* 2KM *of the basin divide. That population density is about constant across portions of the central basin divides where sewer share does not vary suggests that the basin divide does not affect population density independent of sewer share. Figure based on the estimation sample described in column 2 of table 1.*

Appendix B Data

A. Cities

The UN DESA World Urbanization Prospects data (UN DESA Population Division (2018)) is a census of all cities that had a population 300,000 or more in 2014. These data report coordinates for the city center – we frequently refer to this point as the *central business district*. We focus attention on areas that are both within 75KM of the city center and near the boundary of the drainage basin containing the city center.

Intersecting with the census data discussed below, we estimate treatment effects using all cities in the UN Cities data in Brazil, Colombia, South Africa, Jordan, and Tanzania. We can also evaluate counterfactuals in these cities.

B. Sewers

Each of the countries we study provide comprehensive surveys of sewer access at granular geographies in the 2010s. We calculate the share of households in each census geography with sewer access and map the extent of tracts with sewers.

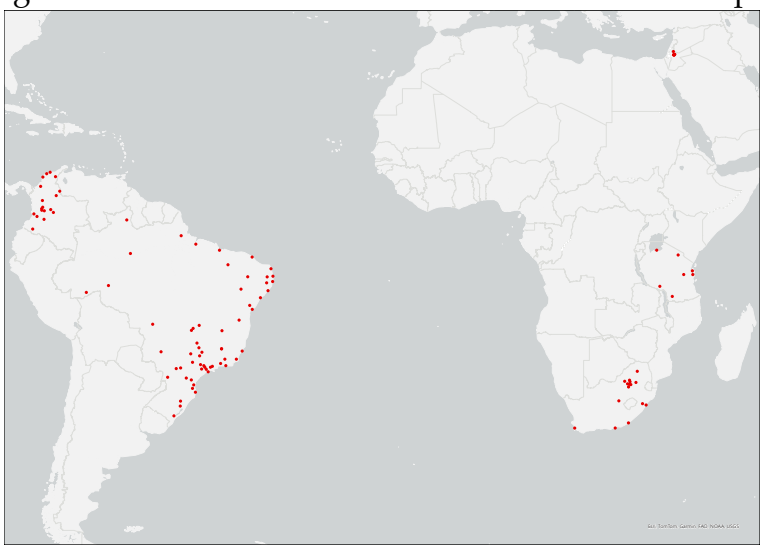
The census in Brazil (Brazilian Institute of Geography and Statistics, 2012) asks households, "Is the bathroom or toilet drain connected to the public sewer system?" The results are reported at the *setores* (English: *sectors*) geographic unit with counts of households affirming and total number of households in the *setor*.

The Colombian census (National Administrative Department of Statistics, 2018) reports counts of households indicating sewer service in response to, "Does your house have sewage service?" This is released at different details of granularity: in urban areas, this is available at the *manzana*-level (English: *block, square*) correseponding to a very fine spatial detail; in rural areas, this is available at the *seccion* (English: *section*) which is larger.

The South African census (Statistics South Africa, 2011) counts households with sewer access using this question: "Is the main type of toilet facility used by this household a flush toilet connected to sewerage system?" The results are reported at a geography the *Small Area Layer* geography, which is between an "enumeration area" and "sub-place."

The census of Jordan (Department of Statistics (Jordan), 2015) asks households, "Does your house have sanitation connected to a public network?"

Figure A2: Locations from UN Cities data in our sample



Note: *Data from the* UN DESA *World Urbanization Prospects (*UN DESA *Population Division, 2018), which is a census of all cities that had a population 300,000 or more in 2014. We consider cities in Brazil, Colombia, South Africa, Jordan and Tanzania.*

The Tanzanian census (National Bureau of Statistics (Tanzania), Office of the Chief Government Statistician, 2012) reports a count of households responding yes to, "Does your house have a flush toilet connected to a piped sewer system?"

Table A4 tabulates more detailed responses on households' sewage disposal for tracts used in the estimation sample from Brazil, Jordan, South Africa, and Tanzania. Most households in the sample were connected to a sewer with cesspits and soak pits being the most common alternative to a sewer connection. Colombia did not provide detailed information on the type of sewage disposal at the manzana level. For Colombian tracts used in the estimation sample, 547,101 households reported being connected to a sewer while 174,009 households were not.

Table A4: Households by sewage disposal type

Country	Sewer	Cess/soak pit	Septic tank	Other/none
Brazil	3,471,439	817,459	592,554	193,767
Jordan	12,273	2,924	0	150
South Africa	468,360	83,229	9,810	28,599
Tanzania	7,636	80,762	15,353	20,614
Total	3,959,708	984,374	617,717	243,130

C. Population density, other outcomes

Using the censuses described in the previous subsection, we calculate population density and literacy measures for all countries, as well as income measures for Brazil and South Africa. For all countries but Jordan, population density is the full count of people divided by tract area. For Jordan, it is the full count of households divided by tract area.

Income is measured monthly in both Brazil and South Africa. Brazil reports the average nominal monthly income of head of household by setore. South Africa reports counts of individuals in 12 income buckets, one of which is "no income." Respondents are asked to consider gross monthly income (pre-tax and including all possible income sources). We approximate the average monthly income for each small area by assigning each income bin the midpoint of the income range for the same bin, then using the reported count to calculate the mean. Colombia, Tanzania, and Jordan do not report income for the granular geographies we use in our analysis.

Literacy is directly reported in the Brazilian and Tanzanian censuses. We do not observe literacy directly in Colombia and South Africa. However, each of these countries releases educational attainment data: Colombia reports a count of persons who have completed *some* amount of primary school (as well as counts for secondary, college/technical, and graduate); South Africa discloses even more granular educational attainment data. We use completion of any primary school, or higher, as a proxy measure for literacy in these countries. We do not observe literacy in Jordan.

D. *Drainage basins*

We construct drainage basins from two digital elevation maps (DEMs) using ArcGIS tools for this purpose. The first DEM derives from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) (NASA/METI/AIST/Japan Spacesystems and US/Japan ASTER Science Team, 2019) and the second from the Shuttle Radar Topography Mission (SRTM) (NASA JPL, 2013). We rely primarily on the ASTER DEM, but consider SRTM for robustness checks.

These data report the elevation of most of the Earth's surface at a spatial resolution of about 30m^2. Using these digital elevation maps, we draw all drainage basins within a 75KM radius of the center of each city using a utility available for this purpose as part of ArcGIS. We identify the drainage basin containing the center of each city using coordinates in the UN DESA World Urbanization Prospects data. These are the central basins, and our research design is organized around comparisons of neighborhoods on opposite sides of the boundaries of these basins.

Figure 3 illustrates basin boundaries for Cascavel, Brazil. This is an empirical analog of the basins drawn in figure 2. In figures 3, black lines indicate the boundaries of drainage basins. Both DEMs are constructed by looking down at the Earth's surface from satellites, and so both sometimes confuse treetops and roofs with ground level. Because ground level elevation is what is relevant for our exercise, this raises the possibility that we mismeasure basin boundaries. ASTER is based on longer wavelength radiation that is better able to penetrate treetops and roofs, and so is a better measure of ground level elevation. Thus, the ASTER data is our primary basis for constructing drainage basin boundaries, and we rely on the SRTM data primarily for robustness checks. A comparison with LIDAR data shows that average error of ASTER is about 4m in four small study areas. SRTM is about the same. (Uuemaa et al., 2020).

Appendix C Small variance approximation of the MTE model

Our data reports tract averages, $(\overline{Y}_j, \overline{D}_j, \overline{W}_j)$, and the variance of W_{ij} within each tract j. Using these data, we write parcel level variables as the sum of tract averages and residuals:

$$W_{ij} = \begin{bmatrix} \overline{X}_j + \sigma\epsilon_{ij}^X \\ \overline{Z}_j + \sigma\epsilon_{ij}^Z \end{bmatrix} = \overline{W}_j + \sigma\epsilon_{ij},$$

where σ^2 is a scaling factor for the residuals, and $\epsilon_{ij} = [\epsilon_{ij}^X, \epsilon_{ij}^Z]$ is a vector of residuals such that the within-tract means satisfy $E(\epsilon_{ij}|j) = 0$ and the tract level variance-covariance matrix for ϵ_{ij}, $Var(\epsilon_{ij}|j) = R_j$ exists. Accordingly, we can express $Var(W_{ij}|j)$ as

$$Var(W_{ij}|j) = \Sigma_{W,j} = \sigma^2 \begin{bmatrix} R_j^{XX} & R_j^{XZ} \\ R_j^{XZ} & R_j^{ZZ} \end{bmatrix}. \qquad \text{(Appendix C.1)}$$

Assume that the parcel level propensity score $p(W_{ij}) = \Pr(D_{ij} = 1|W_{ij})$ is three times continuously differentiable. Then the second-order Taylor expansion of $p(W_{ij})$ with respect to σ around $\sigma = 0$ is,

$$p(W_{ij}) = p(\overline{W}_j) + \nabla p(\overline{W}_j)\sigma\epsilon_{ij} + \frac{\sigma^2}{2}\epsilon_{ij}'\nabla^2 p(\overline{W}_j)\epsilon_{ij} + O(\sigma^3\|\epsilon_{ij}\|^3). \qquad \text{(Appendix C.2)}$$

In this equation, $\nabla p(W_{ij})$ denotes the $1 \times dim(W_{ij})$ vector of partial derivatives of $p(W_{ij})$ with respect to W_{ij} evaluated at W_{ij}, $\nabla^2 p(W_{ij})$ denotes the corresponding Hessian matrix, and $O(\sigma^3\|\epsilon_{ij}\|^3)$ is the residual of the expansion that vanishes as $\sigma^3\|\epsilon_{ij}\|^3 \to 0$.[6] Taking the expectation conditional on i being in the tract j, we have

$$\begin{aligned} p_j \equiv E[D_{ij}|j] = E[p(W_{ij})|j] &= p(\overline{W}_j) + \frac{\sigma^2}{2}E\left[\epsilon_{ij}'\nabla^2 p(\overline{W}_j)\epsilon_{ij}\middle| j\right] + O_j(\sigma^3) \\ &= p(\overline{W}_j) + \frac{\sigma^2}{2}tr\left(\nabla^2 p(\overline{W}_j)\cdot R_j\right) + O_j(\sigma^3), \end{aligned} \qquad \text{(Appendix C.3)}$$

where we let $O_j(\sigma^3) \equiv E[O(\sigma^3\|\epsilon_{ij}\|^3)|j]$, invoking the assumption of finite third-order moments of ϵ_{ij}. The term $\nabla p(\overline{W}_j)\sigma\epsilon_{ij}$ drops out because $E[\epsilon_{ij}|j] = 0$.

We observe $\overline{D}_j$, the share of households in tract j with sewer access. This measures p_j. We also observe $\overline{W}_j$, and $\Sigma_{W,j} = \sigma^2 R_j$. Therefore, given a functional form for the propensity score $p(\cdot)$ and using the tract level observations, we can use (Appendix C.3) to approximately estimate the parcel level propensity score function $p(\cdot)$, where the approximation error is a higher order term of σ than the variance σ^2.

[6] Equation (Appendix C.2) shows the advantage of the representation of $\Sigma_{W,j}$ in equation (Appendix C.1). By introducing the scaling factor σ^2, we facilitate a univariate Taylor series expansion in (Appendix C.2). The same comment applies to equation (Appendix C.5) below.

We now consider the local IV regression (3). Substituting equation (Appendix C.2) into (3) gives,

$$Y_{ij} = X_{ij}^{'}\beta_0 + \left[p(\overline{W}_j) + \nabla p(\overline{W}_j)\sigma\epsilon_{ij} + \frac{\sigma^2}{2}\epsilon_{ij}^{'}\nabla^2 p(\overline{W}_j)\epsilon_{ij} + O(\sigma^3\|\epsilon_{ij}\|^3)\right](\overline{X}_j + \sigma\epsilon_{ij}^X)' \cdot (\beta_1 - \beta_0)$$
$$+ \phi\left(p(\overline{W}_j) + \nabla p(\overline{W}_j)\sigma\epsilon_{ij} + \frac{\sigma^2}{2}\epsilon_{ij}^{'}\nabla^2 p(\overline{W}_j)\epsilon_{ij} + O(\sigma^3\|\epsilon_{ij}\|^3)\right) + U_{ij}. \quad \text{(Appendix C.4)}$$

Next, we take a second-order Taylor expansion of $\phi(\cdot)$, take the conditional expectation given i belonging to tract j as in (Appendix C.3), and finally, absorb higher-order terms in $O_j(\sigma^3)$. This gives,

$$\overline{Y}_j = \overline{X}_j'\beta_0 + p(\overline{W}_j)\overline{X}_j(\beta_1 - \beta_0) + \sigma^2\left[\nabla p(\overline{W}_j)R_j^{WX} + \frac{1}{2}tr\left(\nabla^2 p(\overline{W}_j)\cdot R_j\right)\overline{X}_j'\right](\beta_1 - \beta_0)$$
$$+ \phi(p(\overline{W}_j)) + \phi'(p(\overline{W}_j))\frac{\sigma^2}{2}tr\left(\nabla^2 p(\overline{W}_j)\cdot R_j\right)$$
$$+ \phi''(p(\overline{W}_j))\frac{\sigma^2}{2}\nabla p(\overline{W}_j)R_j(\nabla p(\overline{W}_j))' + O_j(\sigma^3) + \eta_j, \quad \text{(Appendix C.5)}$$

where $R_j^{WX} = E[\epsilon_{ij}\epsilon_{ij}^{X\prime}|j] = \begin{bmatrix} R_j^{XX} \\ R_j^{XZ} \end{bmatrix}$ is a submatrix of R_j.

Finally, assume,

$$\phi(p) = \alpha_0 + \alpha_1 p + \frac{1}{2}\alpha_2 p^2. \quad \text{(Appendix C.6)}$$

Using parameter estimates from equation (Appendix C.3) we can evaluate $p(\overline{W}_j)$ and its derivatives at $\overline{W}_j$. This means that we can define three observable scalars,

$$k_{1j} \equiv \left[\nabla p(\overline{W}_j)R_j^{WX} + \frac{1}{2}tr\left(\nabla^2 p(\overline{W}_j)\cdot R_j\right)\overline{X}_j'\right]$$
$$k_{2j} \equiv tr\left(\nabla^2 p(\overline{W}_j)\cdot R_j\right)$$
$$k_{3j} \equiv \nabla p(\overline{W}_j)R_j(\nabla p(\overline{W}_j))'.$$

Substituting into the structural equation (Appendix C.5), we have

$$\overline{Y}_j = \overline{X}_j'\beta_0 + p(\overline{W}_j)\overline{X}_j(\beta_1 - \beta_0) + \sigma^2(\beta_1 - \beta_0)k_{1j} \quad \text{(Appendix C.7)}$$
$$+ \alpha_0 + \alpha_1\left[p(\overline{W}_j) + \frac{\sigma^2}{2}k_{2j}\right]$$
$$+ \alpha_2\left[\frac{1}{2}p(\overline{W}_j)^2 + p(\overline{W}_j)k_{2j} + \frac{\sigma^2}{2}k_{3j}\right] + O_j(\sigma^3) + \eta_j.$$

We arrive at equation (9) and (10) in the main text by assuming the linear probability model, equation (8), in both (Appendix C.3) and (Appendix C.7), noting $k_{2j} = 0$, and including approximation error in the regression residual.

Our estimand of interest is the SATE. In this case it is given by,

$$\text{SATE} = \overline{X}'(\beta_1 - \beta_0) + \alpha_1 + \frac{1}{2}\alpha_2, \qquad \text{(Appendix C.8)}$$

where $\overline{X}$ is the sample average of X_j. We estimate SATE by plugging in the OLS estimator of the relevant coefficients obtained from (Appendix C.7).

To obtain a standard error estimate for our SATE estimator, we express our estimation of γ and (α,β) using method of moments. Denote the regressor vector (column vector) of equation (Appendix C.7) by,

$$S_j(\gamma) = \left(X_j', \quad (W_j'\gamma)\cdot X_j' + \sigma^2\gamma' R_j^{WX}, \quad W_j'\gamma, \quad \tfrac{1}{2}(W_j'\gamma)^2 + \tfrac{\sigma^2}{2}\gamma' R_j \gamma\right)'$$

and the coefficient vector by $\theta = (\beta_0', \beta_1' - \beta_0', \alpha_1, \alpha_2)'$ with α_0 absorbed into the intercept parameter. Then (θ,γ) is the solution to

$$E\begin{bmatrix} S_j(\gamma)(\overline{Y}_j - S_j(\gamma)'\theta) \\ m_j(\gamma) \end{bmatrix} = 0, \qquad \text{(Appendix C.9)}$$

where $m_j(\gamma)$ is the first-order condition for γ in the OLS estimation of (8). Substituting sample analogs gives our estimators,

$$\frac{1}{n}\sum_{j=1}^{n}\begin{bmatrix} S_j(\widehat{\gamma})(\overline{Y}_j - S_j(\widehat{\gamma})'\hat{\theta}) \\ m_j(\widehat{\gamma}) \end{bmatrix} = 0. \qquad \text{(Appendix C.10)}$$

If we now expand the sample moment conditions around the true parameter values, (θ,γ), we have

$$0 = \frac{1}{n}\sum_{j=1}^{n}\begin{bmatrix} S_j(\gamma)(\overline{Y}_j - S_j(\gamma)'\theta) \\ m_j(\gamma) \end{bmatrix} + \begin{pmatrix} \widehat{\nabla}_{1,\theta} & \widehat{\nabla}_{1,\gamma} \\ O & \widehat{\nabla}_{2,\gamma} \end{pmatrix}\cdot\begin{pmatrix} \hat{\theta}-\theta \\ \widehat{\gamma}-\gamma \end{pmatrix} + \text{remainder}, \qquad \text{(Appendix C.11)}$$

where $\widehat{\nabla}$'s are the derivative matrices of the sample first order conditions.

Multiplying both sides by $\sqrt{n}$, solving for parameters, and letting $n \to \infty$, we obtain the following asymptotic approximation;

$$\begin{aligned}\sqrt{n}\begin{pmatrix} \hat{\theta}-\theta \\ \widehat{\gamma}-\gamma \end{pmatrix} &= -\begin{pmatrix} \nabla_{1,\theta} & \nabla_{1,\gamma} \\ O & \nabla_{2,\gamma} \end{pmatrix}^{-1}\cdot\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\begin{pmatrix} S_j(\gamma)(\overline{Y}_j - S_j(\gamma)'\theta) \\ m_j(\gamma) \end{pmatrix} \\ &\to_d \mathcal{N}(0, \nabla^{-1}\Sigma(\nabla^{-1})'), \qquad \text{(Appendix C.12)}\end{aligned}$$

where

$$\nabla = E\begin{pmatrix} -S_j(\gamma)S_j(\gamma)' & \nabla_\gamma S_j(\gamma)\eta_j - S_j(\gamma)\theta'\nabla_\gamma S_j(\gamma) \\ O & \nabla_\gamma m_j(\gamma) \end{pmatrix} \qquad \text{(Appendix C.13)}$$

$$= E\begin{pmatrix} -S_j(\gamma)S_j(\gamma)' & -S_j(\gamma)\theta'\nabla_\gamma S_j(\gamma) \\ O & \nabla_\gamma m_j(\gamma) \end{pmatrix}. \qquad \text{(Appendix C.14)}$$

The second equality holds because η_j is a regression residual with mean zero conditional on the regressors in (Appendix C.7). Σ is the variance covariance matrix of the moment conditions.

$$\Sigma = E\left[\begin{pmatrix} S_j(\gamma)(\overline{Y}_j - S_j(\gamma)'\theta) \\ m_j(\gamma) \end{pmatrix} \cdot \begin{pmatrix} S_j(\gamma)(\overline{Y}_j - S_j(\gamma)'\theta) \\ m_j(\gamma) \end{pmatrix}'\right]$$

If we specify the linear probability model for (8), then we have $m_j(\gamma) \equiv W_j\nu_j = W_j(\overline{D}_j - W_j'\gamma)$, and $\nabla_\gamma m_j(\gamma) = -W_jW_j'$. Under the same assumption, the derivative matrix of $S_j(\gamma)$ is

$$\nabla_\gamma S_j(\gamma) = \begin{pmatrix} O_{\dim(X)\times\dim(\gamma)} \\ X_jW_j' + \sigma^2(R_j^{WX})' \\ W_j' \\ (W_j'\gamma)W_j' + \sigma^2\gamma' R_j \end{pmatrix}.$$

We estimate the asymptotic variance of (Appendix C.12) by plugging in $\widehat{\gamma}$ in place of γ and replacing the expectation by the sample average for both ∇ and Σ terms,

$$\widehat{\nabla} = \frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} -S_j(\widehat{\gamma})S_j(\widehat{\gamma})' & -S_j(\widehat{\gamma})\hat{\theta}'\nabla_\gamma S_j(\widehat{\gamma}) \\ O & \nabla_\gamma m_j(\widehat{\gamma}) \end{pmatrix},$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} S_j(\widehat{\gamma})(\overline{Y}_j - S_j(\widehat{\gamma})'\hat{\theta}) \\ m_j(\widehat{\gamma}) \end{pmatrix} \cdot \begin{pmatrix} S_j(\widehat{\gamma})(\overline{Y}_j - S_j(\widehat{\gamma})'\hat{\theta}) \\ m_j(\widehat{\gamma}) \end{pmatrix}'$$

Focusing on the first block element of $\widehat{\nabla}^{-1}\widehat{\Sigma}(\nabla^{-1})'$ gives the asymptotic variance estimate for $\sqrt{n}(\hat{\theta} - \theta)$.

Since the SATE estimator can be expressed as $\widehat{\text{ATE}} = a'\hat{\theta}$ with $a = (\mathbf{0}', \overline{X}', 1, 1)'$, its asymptotic variance can be obtained by the asymptotic variance of $\hat{\theta}$ sandwiched by a' and a.