NBER WORKING PAPER SERIES

URBAN SPRAWL AND SOCIAL CAPITAL:
EVIDENCE FROM INDONESIAN CITIES

Andrea Civelli
Arya Gaduh
Alexander D. Rothenberg
Yao Wang

Working Paper 30068
http://www.nber.org/papers/w30068

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2022

We thank seminar participants at Le Moyne College, the 2020 Virtual Meeting of the Urban Economics Association, and Syracuse University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Urban Sprawl and Social Capital: Evidence from Indonesian Cities
Andrea Civelli, Arya Gaduh, Alexander D. Rothenberg, and Yao Wang
NBER Working Paper No. 30068
May 2022
JEL No. D71,H41,R11

ABSTRACT

We use detailed data from Indonesian cities to study how variation in density within urban areas affects social capital. For identification, we instrument density with soil characteristics, and control for community averages of observed characteristics. Under plausible assumptions, these controls address sorting on observables and unobservables. We find that lower density increases trust in neighbors and community participation. We also find that lower density is associated with lower interethnic tolerance, but this relationship is explained by sorting. Heterogeneity analysis suggests that crime in dense areas undermines community trust and participation, intensifying the negative impact of density.

Andrea Civelli
University of Arkansas
Fayetteville, AR 72701
acivelli@uark.edu

Arya Gaduh
Department of Economics
University of Arkansas
402 Business Building
Fayatteville, AR 72701
and NBER
AGaduh@walton.uark.edu

Alexander D. Rothenberg
Syracuse University
426 Eggers Hall
Syracuse, NY 13244-102
adrothen@syr.edu

Yao Wang
Syracuse University
ywang119@syr.edu

1 Introduction

In many developing countries, urbanization proceeds at an astonishing pace. The number of people living in urban areas in Asia increased by more than a billion from 1980 to 2010, and urban populations in Africa are expected to triple between 2018 and 2050 (ADB, 2012; UN, 2019). As rural to urban migration and rising incomes attract millions to cities in lower-to-middle income countries (LMICs), many urban areas are sprawling rapidly. Globally on average, cities are expanding spatially twice as fast as their population growth rates (Angel et al., 2011).

Rapid urban sprawl raises concerns that low-density development may lead to the loss of a sense of community or, more generally, social capital. Critics of urban sprawl have argued that compact, dense urban areas are more likely to promote social interaction (Jacobs, 1961), and because sprawl increases commuting times, it raises opportunity costs for community participation and interaction with neighbors (Putnam, 2000). Moreover, when sprawling development in the urban periphery offers new opportunities for sorting, the clustering of like-minded individuals may also increase political polarization (Bishop, 2009).[1] If lower density erodes social capital, policies to reduce sprawl and mitigate these negative social externalities may be justified (Brueckner and Largey, 2008).[2] Such policies to curb sprawl could include growth controls, binding limits on new construction, or open space dedications (Cunningham, 2007; Glaeser and Ward, 2009; Brueckner and Sridhar, 2012).

However, it is not theoretically clear that low-density development necessarily erodes social capital. For instance, an older tradition in sociology, associated with the works of Simmel (1903) and Wirth (1938), argues that urban life overloads the senses, leading urban residents to limit social interactions with their neighbors to preserve mental energy. Trust may be harder to sustain in denser communities where people from diverse backgrounds and little shared norms interact anonymously (Habyarimana et al., 2007). Moreover, when density is associated with higher crime levels, its negative impact on trust and social capital might become amplified. On the other hand, density could instead facilitate greater intergroup interactions which could foster mutual understanding (Bazzi et al., 2019).

Although correlates of density and social interaction have been well studied in the U.S. and Europe (e.g. Glaeser and Gottlieb, 2006; Brueckner and Largey, 2008), little is known about how low-density development affects social capital in cities from LMICs. In this paper, we revisit this relationship by estimating the causal impact of density on social capital using uniquely rich data from cities in Indonesia, a middle-income and ethnically diverse country that has experienced rapid industrialization. To study multiple dimensions of social capital, we use cross-sectional data from the 2012 National Socioeconomic Survey (*Survei Sosial Ekonomi Nasional*, or *Susenas*). This survey provides detailed measures of various aspects of social capital, including trust in neighbors, community participation, social insurance, and interethnic tolerance. Such measures reflect both "bonding" social capital, which enhances within-group trust, and "bridging" social capital, which strengthens ties across groups (Putnam, 2000). We also use similar measures from a panel of households in the Indonesia Family Life Survey (IFLS).

To estimate the effect of sprawl, we study how changes in density within urban areas impact social

[1] Growing political polarization has been a feature of many LMICs with democratic governments, including Brazil, India, and Indonesia (Carothers and O'Donohue, 2019).

[2] Related theoretical work has also argued that sprawl may reduce social interactions and increase segregation between groups, impacting equilibrium levels of employment (Sato and Zenou, 2015; Picard and Zenou, 2018).

capital outcomes. We confront two fundamental challenges in identifying this relationship: simultaneity and sorting. First, social capital and density may be determined simultaneously within cities, and omitted, place-specific characteristics may drive correlations in both variables. For example, favorable natural amenities may both facilitate cooperation over the provision of local public goods and also attract more people. Second, people who differ in their willingness to contribute to local social capital may sort differentially into areas with different levels of density. For instance, people with a strong dislike of other ethnic groups may sort into more homogeneous areas which tend to be less dense. Both identification problems make it challenging to draw causal inferences from observed correlations between population density and social capital outcomes.

Prior research has confronted the simultaneity problem by instrumenting for population density. For example, Brueckner and Largey (2008) use terrain ruggedness in the urban fringe and population density at a higher level of aggregation to instrument for census tract density. We follow a similar approach, instrumenting density within urban areas using soil characteristics. Soil characteristics were determined millions of years ago, and favorable soils lead to the formation of denser settlements, giving rise to greater density in certain areas of cities that persists today even in the absence of agricultural production. We measure soil characteristics using data from SoilGrids, a global dataset providing high-resolution measures of many soil properties (Hengl et al., 2017). Because we work with a large vector of candidate instruments, many of which may be weak on their own, we follow Belloni et al. (2012) and use post-double-selection lasso techniques to select instruments. This approach obtains the efficiency gains from optimal instruments while reducing problems associated with many instruments.

A key concern with our IV strategy is that even within urban areas, soil characteristics could affect social capital through channels other than density. We provide evidence in favor of the exclusion restriction. First, we show that variation in soil characteristics within cities does not predict contemporaneous agricultural productivity. Our results are robust to excluding both households employed in agriculture and also communities with large shares of agricultural employment. Second, our results are also robust to controlling for historical infrastructure and to controlling for social norms and culture, which could also be influenced by soil characteristics (Alesina et al., 2013). Finally, a placebo exercise using a sample of rural communities where soil characteristics do not predict density largely finds no significant reduced form relationships between the IVs and social capital.

After addressing simultaneity, previous studies of the relationship between sprawl and social capital have not overcome the second identification challenge, namely that sorting could still be confounding estimates. To tackle sorting, we combine instruments for density with controls for community-level averages of observed individual characteristics. We include a rich vector of 38 controls for observed population and demographic characteristics, computed from 2010 census microdata. Altonji and Mansfield (2018) show that under certain assumptions, these controls for sorting on observables will also control for sorting on unobservables. Moreover, we show that by combining their approach—which obtains partial identification of overall group effects—with instruments, we can point-identify the effects of density unconfounded by sorting or simultaneity.

After addressing sorting and simultaneity, we find that conditional on city fixed effects, reductions in density increase trust in neighbors and community participation. Moving from an average suburban neighborhood to one near the central business district (CBD) reduces trust in neighbors by 0.25 standard deviations and community participation by 0.15 standard deviations. These results echo the findings of Brueckner and Largey (2008) who study a similar question in the U.S. We also find that when simultaneity is addressed, lower density is associated with reduced intergroup tolerance and cooperation; however, this result disappears when sorting controls are included. Our results are robust to different specifications, and we find similar results when we use different data to test these hypotheses.

Our finding that greater density reduces within-village trust and community participation contradicts the hypothesis that density affects social capital by reducing the opportunity costs of social interactions (Putnam, 2000; Glaeser and Gottlieb, 2006). A heterogeneity analysis finds that the negative effects of density are also larger for more educated and higher income individuals, and for individuals who commute with a private motor vehicle. We also find that the relationship between density and social capital is more pronounced in higher crime cities. This provides support for the idea that density reduces social capital because crime in dense areas increases fear, undermining trust and community participation (Glaeser and Sacerdote, 1999).

Since the groundbreaking work of Putnam (2000), a large empirical literature spanning multiple disciplines tries to estimate the impact of urban sprawl on social capital, but existing evidence is mixed. Many studies find a negative relationship between various aspects of sprawl and social capital, but these results are difficult to interpret because they do not adequately address challenges of causal inference.[3] More recent studies have attempted to address the endogeneity of sprawl measures, such as density, through the use of instrumental variables or other methods. Although these studies do not tackle sorting, they tend to find that social capital is lower in higher-density communities.[4] Our estimates, which use both instruments for density and controls for sorting, represent a methodological improvement over prior research.

We also contribute to the scarce literature studying the relationship between sprawl and social capital in LMIC cities. With worse traffic congestion and lower-quality public transport infrastructure, the potential for sprawl to impact social capital in urban communities in LMICs could be even larger than in high-income countries. The few existing empirical papers on sprawl and social capital in LMICs, including Hemani et al. (2017) on neighborhood forms and social capital in Assam, India, and Zhao (2013) on segregation and sprawl in Beijing, suffer from identification problems. Another body of research uses qualitative methods, such as Connell (1999) on social isolation and sprawl in Manila, Caldeira (2001) on social segregation and fear in Sao Paolo, and Coy and Pöhler (2002) on gated communities in Latin American cities. Our paper represents some of the best quantitative evidence on sprawl and social capital from a non-upper income context.[5]

Methodologically, our study illustrates one of the first applications of the framework of Altonji and Mansfield (2018) to identify the causal effect of a particular group-level treatment on individual outcomes in the presence of sorting into groups. This strategy of combining administrative data to construct

[3] Prominent examples include Freeman (2001) on U.S. cities in the early 1990s, Leyden (2003) on the effects of walkable neighborhoods in U.S. cities, Besser et al. (2008) on commuting times and socially oriented trips, Wood et al. (2008) and Wood et al. (2012) on the impact of neighborhood design and social interactions in Western Australia, and Glaeser and Sacerdote (2000) on the impact of living in large apartment buildings.

[4] Examples of studies focusing on the United States and using instruments for density or other approaches to address the simultaneity problem include Glaeser and Gottlieb (2006), Brueckner and Largey (2008), and Nguyen (2010).

[5] In a recent paper, Muzayanah et al. (2020) also study sprawl and social capital outcomes in Indonesian cities using similar data as we do, but our methodologies differ substantially. As we do, they find that individuals living in higher-density areas had lower levels of trust, but they do not address any endogeneity issues.

controls for sorting with instruments for particular group-level treatments can be broadly applied, for instance, in studying the impact of specific educational interventions or the effect of certain neighborhood-level treatments.

The rest of this paper is organized as follows. Section 2 presents background information on urbanization, economic development, and social capital in Indonesia. Section 3 describes the different datasets we analyze. Section 4 explains how we define metropolitan areas, how we measure urban sprawl, and presents some evidence on correlates of sprawl. Section 5 describes our empirical strategy. Section 6 presents our results, and section 7 concludes.

2 Background: Indonesia's Urban Transformation

After a sustained period of economic growth and structural transformation since the late 1960s, Indonesia is now a lower-middle income economy. From 1970 to 1997 (immediately before the Asian Financial Crisis), per-capita GDP grew by 4.4 percent a year, from just $772 in 1970 (in constant 2010 USD) to just over $2,400 in 1997. This sustained growth was accompanied by rapid structural change as agriculture's share of GDP rapidly declined, while the share of manufacturing and services grew dramatically (Hill, 1996).

As the structure of Indonesia's economy transformed, people increasingly migrated away from rural areas, leading to rapid urbanization. In the 1980s and 1990s, the country's urbanization rate grew by 3 percent a year—faster than many other East Asian countries, including China. Between 1990 and 2000, the rate of urban growth peaked, and the subsequent two decades saw much slower urbanization. The population living in urban areas more than doubled between 1970 and 2010, and today, about 151 million Indonesians live in urban areas (56 percent of the population Roberts et al., 2019). By 2045, when Indonesia will celebrate its centennial, cities are expected to house roughly 220 million people (70 percent of the population).

As urban growth continues, many Indonesian cities have experienced significant spatial expansion. Economic activities in the largest cities have sprawled well beyond their administrative borders, leading to the formation of agglomerations that span multiple districts.[6] Although the high economic productivity in urban cores attracted significant migration and growth, housing costs increased rapidly, and many urban residents relocated to the periphery. Indeed, between 2000 and 2010, about one-third of population growth in the peripheries of metro areas has come from migration (Roberts et al., 2019, Figure 1.12).

Although variation in the extent of urban sprawl across cities is driven by geo-climatic and socio-economic characteristics, national development policies also played an important role (Civelli and Gaduh, 2018). For example, Indonesian leaders have generally enacted policies favoring motor vehicles, including subsidizing gasoline prices and investing in road construction, instead of making public transport investments. The agencies responsible for managing land use have generally been ineffective in using policy levers, including binding limits on new construction, open space dedications, growth controls, or environmental regulations, that could potentially limit sprawl (Rukmana, 2015).[7]

[6] Roberts et al. (2019, Box 1.5) identified a total of 21 multidistrict metro areas, defined as metro areas whose labor markets span multiple districts.

[7] Rukmana (2015) also finds that only 8 percent of the land permitted for housing in West Java was complaint with spatial plans.

Social Capital and Diversity. Following Putnam (1995), we define social capital as the features of social life, including social networks, social norms, and trust, that enable community members to act together effectively to pursue shared objectives.[8] Using a set of indicators that capture this notion of social capital, Legatum Institute (2019a) ranked Indonesia 5th out of 167 countries.[9] Indonesia received this high ranking despite being one of the world's most diverse countries, with more than 1,200 self-identified ethnic groups whose members belong to one of six recognized religions.[10] In 2010, its national-level ethnic fractionalization index—the probability that any two residents, randomly drawn from the national population, belong to different ethnicities—was around 0.81.

Indonesia's high ranking on the social capital index amidst its national diversity may appear to contradict an extensive empirical literature documenting a negative association between ethnic diversity and social capital (e.g. Alesina and La Ferrara, 2002, 2000; Costa and Kahn, 2003; Putnam, 2007). However, these contradictions disappear once we look at the local level. First, despite its national diversity, most Indonesian communities are very homogeneous, explained in part by the nation's archipelagic geography, which separates ethnic groups by vast waterways. The median community has a very low ethnic fractionalization index of 0.04.[11] Second, the negative association between diversity and social capital resurfaces when we look at variation within Indonesia (Mavridis, 2015; Gaduh, 2016). Ethnic and religious tensions between groups have occasionally sparked violent conflicts throughout Indonesian history (Bertrand, 2004) and such conflicts are more likely when different groups are residentially clustered (Barron et al., 2009). Finally, the extent to which local diversity negatively affects social capital depends on whether diversity makes local inter-group competition more salient (Bazzi et al., 2019).

A variety of socio-cultural institutions foster social interactions at the local level, particularly in rural areas. For example, "*gotong royong*" is a norm of mutual and reciprocal assistance rooted in village, agrarian societies across Indonesia (Bowen, 1986). It encourages community participation through mutual insurance, public goods provision, collective work (e.g., harvesting), as well as births, weddings, and funerals (Koentjaraningrat, 1985). Many of these social activities are organized locally by social, religious, or village-government organizations. Others, such as mutual insurance through *arisan* (i.e., rotating savings and credit associations, or ROSCA), were often developed organically by individual members.

Local-level institutions, and the social capital they generate, can help to fill some of the gaps in public goods left by the Indonesian government, both in rural and urban communities (Woolcock and Narayan, 2000; Bebbington et al., 2006). Many ethnic and religious institutions can naturally be found in both rural and urban communities. However, as sprawl expands cities into previously rural communities, the melding of rural and urban communities — what McGee (1991) describes as "*desakota*" (village-city)

He argued that spatial planning was used to accommodate new construction and benefit real estate developers connected to Suharto's former regime, rather than to control undesirable development.

[8] This definition is consistent with DiPasquale and Glaeser (1999), who argue that while social capital does not directly enter the utility function, it enhances the ability of neighbors to enjoy private investments in community organizations, social groups, or local public goods.

[9] The Legatum Institute uses survey responses to construct a country-level social capital index that captures "the personal and family relationships, social networks and the cohesion a society experiences when there is high institutional trust, and people respect and engage with one another" (Legatum Institute, 2019b, p.7).

[10] The Indonesian government officially recognizes only six religions: (1) Buddhism; (2) Catholicism; (3) Confucianism; (4) Hinduism; (5) Islam; and (6) Protestantism.

[11] Formally, the ethnolinguistic fractionalization index for community c can be written as $ELF_c \equiv 1 - \sum_g \pi_{g,c}^2$, where $\pi_{g,c}$ is the share of group g in the population of community c.

— means that some of the social institutions originating in rural areas (e.g., collective maintenance of public goods, mutual insurance) can also be found in urban peripheries (Beard and Dasgupta, 2006). There is also evidence that social interactions remain just as vibrant in these urban communities as they do in rural areas (Jellinek, 1991; Wilhelm, 2011).

3 Data

To study the relationship between social capital and urban sprawl in Indonesia, we combine several high-quality data sources. These include social capital measures from household survey data, population census data, and geospatial datasets. We briefly describe our main data sources here, introducing others as we use them in the analysis.

Social Capital Measures. Our primary source of social capital outcomes is the 2012 National Socioeconomic Survey (*Survei Sosial Ekonomi Nasional*, or *Susenas*). The 2012 *Susenas* contains a detailed module that asks household-head respondents several questions about different aspects of social capital. These variables, described and summarized in Table 1, have been grouped into four broad categories. Panel A lists questions that ask how well individuals trust their immediate neighbors, an important dimension of local social capital. Panel B contains questions that measure participation in community activities. Panel C lists questions that refer to social insurance, namely the extent to which individuals are willing to assist or expect to receive help from their neighbors in times of financial hardship or natural disasters. Panel D includes measures of tolerance of other ethnic and religious groups.

Our main specifications include responses to these questions for roughly 20,000 households living over 2,200 communities (*desa* or *kelurahan*). The community is the lowest administrative unit in Indonesia and comprises our main spatial unit of analysis.[12] The communities we study are spread throughout cities in Indonesia, spanning 27 of Indonesia's 34 provinces. Our main analysis focuses on estimates of the mean effects of density on these groups of outcomes, following Kling et al. (2007) as we describe below. In addition, the *Susenas* data also contain measures useful for individual-level controls, including age, education, marital status, sector of employment, and employment status.

Community-Level Demographic Characteristics. We construct community-level demographic characteristics with 2000 and 2010 Population Census data. These data allow us to construct multiple measures, including population density at the community level, the share of community members with different levels of educational attainment, and the share of the population that is married or migrated from another district. The data also include questions on ethnicity, religion, and census block information that we use to construct diversity and segregation measures. As we describe below, community averages of individual-level characteristics, which we calculate with these data, are crucial for our empirical strategy.

Soil Characteristics. We use data from SoilGrids to measure the characteristics of soils prevalent in Indonesian communities. SoilGrids is a global dataset combining hand-collected soil profiles from nearly 150,000 sites with machine learning algorithms to provide global, 250-meter resolution predictions of many standard soil properties (Hengl et al., 2017).[13] These properties include (1) bulk density; (2) water

[12] According to Census data, the communities in our sample had an average population of 9,766 in 2000 and 11,462 in 2010.

[13] Hengl et al. (2017) harmonize characteristics from soil samples collected across all 7 continents and multiple countries, and

content; (3) sand content; (4) clay content; (5) texture classification; and (6) soil taxonomy information. Although measures of organic carbon content and soil pH are also available, we did not use these measures in our analysis, because they can be directly manipulated by human activity. We also only used soil characteristics measured at a depth of 60 cm or more, as these capture variation in the subsoils and parent material of soils which were largely determined millions of years ago.

Global Human Settlements Data. To measure changes in the built-up extent of urban areas and to define our urban sample, we rely on data from the Global Human Settlement Layer (GHSL), produced by the European Commission's Joint Research Centre (JRC). These data were created by applying machine learning techniques to 40 years of Landsat satellite imagery to measure the locations of human settlements, including buildings and physical infrastructure (Pesaresi et al., 2016). To calculate sprawl measures for Indonesian cities and to measure the spatial extent of urban areas, we use GHS-BUILT grids from 1990, 2000, and 2014. These data report the share of each 30-meter pixel that is classified as containing built-up surfaces.[14]

Geospatial Data on Administrative Boundaries and Topography. Our analysis also relies on administrative boundary shapefiles that identify community borders. These datasets are created by Indonesia's national statistical agency, *Badan Pusat Statistik* (BPS). We use these boundaries in combination with data from the Harmonized World Soil Database (HWSD) to construct basic topographic characteristics (e.g., ruggedness, slope, and elevation).

4 Measuring Urban Areas and Sprawl

To study the relationship between social capital and urban sprawl, we focus only on communities that comprise Indonesia's metropolitan areas. Although BPS provides rural and urban definitions, such measures often reflect political boundaries and do not capture the full economic borders of cities. In the absence of reliable government definitions, there is no unambiguous way to classify urban areas and determine which areas are part of which cities. This is a notoriously difficult problem, and multiple approaches for classifying areas as urban and assigning them to different cities have been suggested in the literature (e.g. Chomitz et al., 2005; Uchida and Nelson, 2010; Duranton, 2015).[15]

We adopt a morphological approach to city definitions (see Burchfield et al., 2006). We begin by identifying cities using a list of 83 urban regions in Indonesia from the World Bank's East Asia and Pacific Urban Expansion (EAP-UE) project. This list contains administrative areas with populations of 100,000 or more in 2010. We then carve out the physical boundaries of each city from these EAP-UE urban regions. To do so, we identify the borders of urban areas based on the locations of built-up surfaces in 2000, as measured with 30-meter resolution GHS-BUILT data. For any built-up pixel,

they train machine learning algorithms to predict those characteristics using covariates derived from remote sensing data. The dataset they produce is publicly available through Google Earth Engine's Data Catalog.

[14] Note that when measuring human settlements, we only use the GHS-BUILT grids from GHSL, and we refer to GHS-BUILT and GHSL interchangeably. For our sample of urban areas in Indonesia, most of the input data for the 1990 GHSL come from Landsat tiles collected in 1989-1990 (as shown in Gutman et al., 2013, Figure 3). More recent GHSL data are covered by annual Landsat data. We do not work with the 1975 GHS-BUILT because of coverage gaps. As shown in Appendix Figure A.1, large portions of Indonesia, including the entire island of Sumatra and portions of West Java, are missing in the 1975 epoch.

[15] Bosker et al. (2021) discuss how different approaches may be used to define urban areas in Indonesia.

its urban development density is defined as the percentage of built-up space in the immediate square kilometer surrounding it.

We classify a city's *core area* as consisting of those built-up pixels that lie within the administrative boundaries of an EAP-UE urban region and are surrounded by land that is more than 50 percent built-up. Typically, our definition of an urban core identifies a large, compact block of pre-existing built-up areas that correspond to the inner part of a city. However, smaller satellite centers that satisfy the core classification criteria might also exist around the main core.

Around this high-density core and within the administrative boundaries of the metropolitan region, we also define the *urban fringe* of the city—where urban spatial expansion occurs between earlier years and 2014—using a 20-kilometer buffer area.[16] Appendix Figures A.2 and A.3 provide an illustration of the core-fringe identification for the metropolitan area of Bandar Lampung in South Sumatra, which includes both the city (*kota*) of Bandar Lampung and the surrounding district (*kabupaten*) of Lampung Selatan. It is worth noting that in this procedure, the core of the metropolitan area does not necessarily match the administrative boundaries of the *kota*. Similarly, the metropolitan area is smaller than the simple union of the two administrative units.

This approach identifies 80 urban metropolitan areas in Indonesia out of the 83 metropolitan areas initially listed by the EAP-UE project. The remaining 3 areas were dropped because they either lacked a well-identified core or did not exhibit sufficiently strong urban expansion in 2014.[17] Figure 1 illustrates the geographic distribution of these 80 metropolitan areas. Half of these areas are located on the Inner Islands of Java and Bali, a quarter are on Sumatra, and the remaining quarter are in other parts of the Outer Islands. The largest metropolitan area is Greater Jakarta (*Jabodetabekpunjur*), the economic and political center of Indonesia, which is a megacity of over 30 million people. Three other cities have more than 2 million inhabitants—these are Bandung, Surabaya, and Medan—while others have between 100,000 and 2 million people.[18]

In our analysis below, we only include communities in our sample if they are part of at least one metro area, based on our definitions. A total of 20,717 communities are in our sample (out of 75,267 total in Indonesia), but only 2,288 communities from 76 metro areas were covered by the 2012 *Susenas*.[19]

Measuring Sprawl and its Correlates. According to GHSL data, 0.46 percent of Indonesia consisted of built-up areas in 1990. By 2014, that figure had nearly doubled to 0.75 percent. To measure the extent of urban sprawl for cities in Indonesia, we follow Burchfield et al. (2006). For each 30-meter newly built-up cell in our urban areas in 2014 (relative to 2000, the previous epoch), we calculated the share of open space in the immediate square kilometer. The sprawl index for the urban area is the average of those open space measures for these pixels. It provides a direct measure of undeveloped land in the square

[16] Burchfield et al. (2006) choose a 20-kilometer fringe because it contains almost 100 percent of the new developments around built-up areas in the U.S. at the beginning of their sample. Visual inspection of maps produced in our analysis confirms that this also holds for Indonesia.

[17] The cities that were dropped include: (1) Bontang (East Kalimantan); (2) Maluku (Ambon); and (3) Jayapura (Papua).

[18] The concentration of metropolitan areas in the Inner Islands closely reflects differences in economic and urban development across the regions of the archipelago. The Inner Islands contain 60 percent of Indonesia's population but only 8 percent of the country's land area. Modern economic activity has always been concentrated in the Inner Islands, and economic activities there contributed about two-thirds of the national GDP in 2004 (Hill et al., 2009). While manufacturing and services are concentrated on the Inner islands, agriculture is still predominant in the Outer Islands.

[19] Note that our definitions of urban areas are somewhat different from BPS classifications. For instance, 671 communities (29.3%) in our urban *Susenas* sample are classified as rural communities by BPS.

kilometer surrounding an average residential development.

Figure 2 illustrates significant variation in sprawl across metropolitan areas. The sprawl index ranges from a minimum of 65.8 to a maximum of 92.3, with a mean of 80.3. Metropolitan areas in the Outer Islands, where the share of urban land cover is still quite low and cities can expand into more undeveloped areas, are typically less compact. The index is generally lower for major cities like Jakarta and Bandung, as one might expect. Sprawl in Indonesian cities is also much larger than sprawl in major U.S. cities; the mean sprawl index for cities with populations of 1 million or more in the U.S. was only 38.9, according to Burchfield et al. (2006, Table 2).

Urban sprawl is neither a recent phenomenon for Indonesia's metro areas, nor has it been uniform over time. Using GHS-BUILT data for 1990, we also construct a sprawl index for urban areas between 1990 and 2000, and compare it to the sprawl index between 2000 and 2014. Figure 3 presents a scatter plot of the indexes for these two periods relative to the 45-degree line, where cities would fall if they experienced the same amount of sprawl over both periods. While sprawl is positively correlated between the two periods, the pace of sprawl increased in the 2000s, as most cities lie above the 45-degree line. The effect is also most pronounced for cities with lower sprawl in the 1990-2000 period.

As cities sprawl, density falls, and new housing constructed in the periphery offers increased opportunities for sorting. Figure 4 plots the relationship between a community's distance to the central business district (CBD) and several community-level variables measured from 2010 Census data.[20] The figure only uses data for communities that comprise the cores and peripheries of urban areas, our main analysis sample. Estimated local polynomial regression lines for the relationships are reported in red, along with confidence bands in gray. Panel A shows that population density declines substantially as distance to the CBD increases. Panel B plots the relationship between community-level ethnic fractionalization and distance to the CBD. This figure shows that communities located farther from the center of the city are more ethnically homogeneous. Ethnic fractionalization is highest in the cores of metropolitan areas, but it displays a sharp decline of about 50 percent in the first 10 km from the CBD, flattening out after that. Panel C shows that religious fractionalization declines similarly as distance to the CBD increases, tapering off again after a distance of 10 km. Panel D plots the relationship between city-level sprawl and ethnic segregation, using the Alesina and Zhuravskaya (2011) segregation measure applied to communities in each urban area.[21] Ethnic segregation increases moderately as sprawl increases, suggesting that sprawling cities may provide more opportunities for sorting by ethnicity.

[20] Distance to the CBD is defined as the crow-flies distance between the centroid of the urban core polygon and the community's centroid, measured in kilometers.

[21] Let c index cities, let $v = 1, ..., V^c$ index communities within city c, and let $k = 1, ..., K$ index ethnic groups. The Alesina and Zhuravskaya (2011) segregation measure, a squared coefficient of variation between community ethnic group shares and the shares of ethnic groups in the city's population, is defined as follows:

$$S_c = \frac{1}{K-1}\sum_{k=1}^{K}\sum_{v=1}^{V^c}\frac{n_v}{N_c}\frac{(\pi_{k,v}-\pi_k^c)^2}{\pi_k^c}$$

where n_v is community v's population, N_c is the population of city c, $\pi_{k,v}$ is the share of group k in community v, and π_k^c is the share of group k in city c's population. If each community in city c were comprised of a separate group, S_c would equal 1, reflecting full segregation. If each community in city c had ethnic group shares that were equal to the city's overall ethnic shares, S_c would equal zero, reflecting perfect integration.

5 Empirical Strategy

In this section, we explain our approach for addressing the two key identification challenges that confound estimates of the relationship between density and social capital: (1) sorting of individuals with lower or higher costs to contributing to social capital; and (2) the simultaneous determination of density and social capital by unobserved place-specific variables. Our empirical strategy builds on the control function approach of Altonji and Mansfield (2018) for bounding the variance of overall group treatment effects in the presence of sorting into groups, but we add instruments to point identify the effect of one particular group attribute. We describe key features of the procedure here but leave many details for Appendix C.

5.1 Sorting into Communities

Let i index individuals and let $v \in \{1, ..., V\}$ index the discrete set of communities comprising different metropolitan areas in Indonesia. Individual i's consumer surplus from choosing to live in community v is given by the following expression:

$$U_i\left(v\right) = \mathbf{W}_i \mathbf{A}_v - P_v + \varepsilon_{iv}\,, \tag{1}$$

where $\mathbf{A}_v$ represents a $(K \times 1)$ vector of amenities that characterize community v, P_v is the price of living in community v, and ε_{iv} is an idiosyncratic component specific to individual i's tastes for living in community v. The term $\mathbf{W}_i$ represents a $(1 \times K)$ vector of weights measuring i's willingness to pay for different components of the amenity vector. Note that $\mathbf{A}_v$ could contain endogenous amenities, such as density, which are determined in equilibrium by the sorting process.

We partition $\mathbf{W}_i$ into three components: (1) $\mathbf{X}_i$, a vector of individual-level observables that influence tastes for amenities and social capital outcomes; (2) $\mathbf{X}_i^U$, a vector of individual-level unobservables that influence tastes for amenities and social capital outcomes; and (3) $\mathbf{Q}_i$, a vector of variables (both observed and unobserved) that may influence preferences over amenities and sorting but have no impact on social capital outcomes:

$$\mathbf{W}_i = \mathbf{X}_i \mathbf{\Theta} + \mathbf{X}_i^U \mathbf{\Theta}^U + \mathbf{Q}_i \mathbf{\Theta}^Q\,,$$

where $\mathbf{\Theta}$, $\mathbf{\Theta}^U$, and $\mathbf{\Theta}^Q$ are the respective willingness to pay coefficients. Note that we define $\mathbf{X}_i$ and $\mathbf{X}_i^U$ so that they represent the complete set of individual factors that determine social capital outcomes. As emphasized by Altonji and Mansfield (2018), this formulation allows for a fairly general pattern of relationships between individual characteristics (both observable and unobservable) and tastes for amenities, subject to the assumption that the indirect utility function is additively separable, as expressed in equation (1).

We assume that individuals take prices, P_v, and amenities, $\mathbf{A}_v$, as given when making location decisions, and that individuals choose the community that maximizes (1) using all information available to them.[22] This information set includes housing prices in different locations, the vectors of amenities in

[22] If population density is an element of $\mathbf{A}_v$, we can assume that agents form expectations about the density that will prevail in each community before they move. When they move, because agents act atomistically, they ignore their impact on the equilibrium density that emerges.

those locations, the full set of preference weights, $\mathbf{W}_i$, and realizations of the idiosyncratic component, ε_{iv} for all $v \in \{1, ..., V\}$. Let $v(i)$ denote the optimal community choice for individual i.

Altonji and Mansfield (2018) prove that given this setup and under a relatively weak set of additional assumptions, the community-level expectation of individual-level unobservables that influence social capital, denoted by $\mathbf{X}_v^U \equiv \mathbb{E}[\,\mathbf{X}_i^U \,|\, v(i) = v\,]$, is linearly dependent on community-level average observables, $\mathbf{X}_v \equiv \mathbb{E}[\,\mathbf{X}_i \,|\, v(i) = v\,]$. The intuition behind this argument is that sorting creates two vector-valued mappings: (1) a mapping between community-level averages of observables and amenities in that community, denoted by $\mathbf{X}_v = \mathbf{f}\left(\mathbf{A}_v\right)$; and (2) a mapping between community-level averages of unobservables in community v and amenities, denoted by $\mathbf{X}_v^U = \mathbf{f}^U\left(\mathbf{A}_v\right)$. The authors provide conditions under which the first mapping, $\mathbf{f}$, is invertible, so we can write: $\mathbf{X}_v^U = \mathbf{f}^U\left(\mathbf{f}^{-1}\left(\mathbf{X}_v\right)\right)$. Under an additional assumption, the relationship between $\mathbf{X}_v^U$ and $\mathbf{X}_v$ induced by composing these vector-valued functions is actually linear.[23]

The strongest of these assumptions is the spanning assumption (assumption A5 in Altonji and Mansfield, 2018) which states that the coefficient vectors $\mathbf{\Theta}^U$, which relate tastes for amenities to elements of $\mathbf{X}_i^U$, need to be linear combinations of $\mathbf{\Theta}$, which relate tastes for amenities to elements of $\mathbf{X}_i$ and/or elements of $\mathbf{X}_i^U$ that are correlated with $\mathbf{X}_i$. One of the two sufficient conditions for this spanning assumption to hold is that $\mathbf{f}$ is invertible. A necessary condition for invertibility is that the dimension of $\mathbf{A}^{\mathbf{X}}$, the subset of amenities that affect the distribution of community averages, is less than the number of elements in $\mathbf{X}_v$. This would occur if $\mathbb{V}\left(\mathbf{X}_v\right)$ is rank deficient.

These results suggest that with a rich enough dataset, we can use community-level averages of individual-level observable characteristics to effectively control for sorting. In our empirical implementation, we use a vector of 38 variables constructed from unit-level 2010 census data to measure $\mathbf{X}_v$. These variables include the community's average age, years of schooling, household size, the percentage of the community that is female, the percent who self-identify with different religions or ethnicities, the share of different types of employment status and marital status, and the share who speak Indonesian at home.[24] Appendix Table A.1 reports a principal components analysis of these 38 $\mathbf{X}_v$ variables. In our urban *Susenas* sample (column 2), only 27 factors explain 95 percent of the total variation in $\mathbf{X}_v$, 32 factors explain 99 percent of the total variation in $\mathbf{X}_v$, and 37 factors explain 100 percent of the total variation in $\mathbf{X}_v$. This suggests that for the urban *Susenas* sample, $\mathbf{X}_v$ is rank deficient.

Appendix Table A.2 also formally tests hypotheses about the rank of the $\mathbf{X}_v$ covariance matrix, using a test proposed by Kleibergen and Paap (2006). We find that for the full *Susenas* sample, we cannot reject the null hypothesis that the rank of the variance-covariance matrix of $\mathbf{X}_v$ is 34 against the alternative that it is 35 or greater. For the urban sample, we cannot reject the null hypothesis that the rank of the variance-covariance matrix of $\mathbf{X}_v$ is 28 against the alternative that it is 29 or greater. The results from Appendix Tables A.1 and A.2 suggest that because $\mathbf{X}_v$ is rank deficient, $\mathbf{f}$ will be invertible, so that $\mathbf{X}_v$ can be used as a linear control function for sorting on unobservables.

[23] The full set of assumptions is explained in more detail in Appendix C.

[24] If the 2010 census data are inaccurate, they could provide potentially noisy measures of $\mathbf{X}_v$, the expected values of observable characteristics in community v. However, Altonji and Mansfield (2018) provide a Monte Carlo analysis suggesting that even with small samples from survey data (i.e. $N = 20$), we can approximate $\mathbf{X}_v$ fairly well.

5.2 Production of Social Capital

After individuals choose locations, we assume that a social capital outcome for individual i living in community v, denoted by y_{vi}, is produced according to the following linear, additively separable function:

$$y_{vi} = \mathbf{X}_i\beta + x_i^U + \theta \log \text{density}_v + \mathbf{C}_v\mathbf{\Gamma} + c_v^U + \eta_{vi} + \xi_{vi} \,. \tag{2}$$

Because many outcomes recorded in the 2012 *Susenas* data are either binary or take on discrete values (often 4-point scales), y_{vi} is the continuous latent variable that determines these values. Equation (2) is composed of three sets of terms: (1) an individual component; (2) a community-level component; and (3) an idiosyncratic component. We describe each of these components in detail.

The individual component, $\mathbf{X}_i\beta + x_i^U$, includes a row vector, $\mathbf{X}_i$, collecting individual i's observed attributes, and the parameter β measures how those attributes affect y_{vi}. The second part consists of a scalar, $x_i^U \equiv \mathbf{X}_i^U\beta^U$, which summarizes the contribution of unobserved individual characteristics ($\mathbf{X}_i^U$) to social capital outcomes.

The community-level component, $\theta \log \text{density}_v + \mathbf{C}_v\mathbf{\Gamma} + c_v^U$, contains three terms. The first measures log population density at the community level, where density is defined as the population of community v in 2010 divided by the area of that community (in square km). The key object of interest, θ, measures the semi-elasticity of social capital outcomes with respect to density. The second component is a row vector, $\mathbf{C}_v$, capturing the influence of other observed community-level characteristics on social capital outcomes. We include urban-area fixed effects in $\mathbf{C}_v$. Finally, the third term, $c_v^U \equiv \mathbf{C}_v^U\mathbf{\Gamma}^U$, represents a scalar that summarizes the contribution of unobserved neighborhood characteristics to y_{vi}.

Finally, the idiosyncratic component, $\eta_{vi} + \xi_{vi}$, also contains two terms. The first term, η_{vi}, captures unobserved variation in community contributions to social capital among individuals who live in that community. Some factors correlated with η_{vi} may be captured by observed and unobserved community-level variables. The second term, ξ_{vi}, captures other influences to y_{vi} that are determined after that individual arrives in community v, but are unpredictable given $\mathbf{X}_i$, x_i^U, $\log \text{density}_v$, $\mathbf{C}_v$, c_v^U, and η_{vi}. Such influences could include local labor market shocks that make it harder or easier to participate in the community, or shocks to local public goods that influence individuals differently in certain areas.

We partition the group-level observables (excluding log density) into $\mathbf{C}_v = [\mathbf{X}_v, \mathbf{C}_{2v}]$, and we partition their coefficients analogously, so that $\mathbf{\Gamma} = [\mathbf{\Gamma}_1, \mathbf{\Gamma}_2]$. The term $\mathbf{X}_v$ includes community averages of individual-level observables (our sorting controls), while the term $\mathbf{C}_{2v}$ includes community-level characteristics that are not mechanically related to community composition. In our baseline specifications, these include pre-determined, exogenous natural amenities, such as elevation, ruggedness, and distance to the coast or rivers, which may make it easier or harder to sustain a social capital outcome. This notation lets us we rewrite equation (2) as follows:

$$y_{vi} = \mathbf{X}_i\beta + x_i^U + \theta \log \text{density}_v + \mathbf{X}_v\mathbf{\Gamma}_1 + \mathbf{C}_{2v}\mathbf{\Gamma}_2 + c_v^U + \eta_{vi} + \xi_{vi} \,. \tag{3}$$

Note that because of the assumptions described in Section 5.1 above, adding $\mathbf{X}_v$ effectively controls for sorting on both observables and unobservables in the production of social capital. Although a typical control function procedure would use a non-linear or semi-parametric control function, the spanning assumption (assumption A5 in Altonji and Mansfield, 2018) implies that we just need to include these controls linearly.

5.3 An Instrumental Variables Estimator for θ

Altonji and Mansfield (2018) use $\mathbf{X}_v$ controls to partially identify the contribution of total group treatment effects (e.g., school or neighborhood effects) to outcomes. When estimating this overall group treatment effect, controlling for group averages eliminates the sorting bias, but it may also over-control, because peer effects may depend on these group-averages.[25] Consequently, they can only obtain a lower bound on the overall importance of school or neighborhood effects in explaining the variance in outcomes. We extend their approach by introducing an instrument for a particular group attribute (namely density) to point-identify the effect of that attribute on outcomes in a way that is unconfounded by sorting.[26]

Let $\widetilde{\mathbf{X}}_{iv} \equiv [\mathbf{X}_i, \mathbf{X}_v, \mathbf{C}_{2v}]$ collect the observed variables that do not include log density, and let $\widetilde{\beta} = [\beta, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2]$ collect their parameters. Also let $u_{iv} \equiv x_i^U + c_v^U + \eta_{vi} + \xi_{vi}$ collect all of the unobserved components. Using this notation, we can simplify (3) even further:

$$y_{vi} = \theta \log \text{density}_v + \widetilde{\mathbf{X}}_{iv}\widetilde{\beta} + u_{iv}\,.$$

Let $\mathbf{Z}$ denote a vector of instruments for density and let $\widetilde{\mathbf{X}}_{iv}$ act as instruments for themselves. An IV estimator for θ can be written as:

$$\widehat{\theta}_{IV} = \left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}} \log \text{density}\right)^{-1} \mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\mathbf{y}\,, \tag{4}$$

where $\mathbf{M}_{\widetilde{\mathbf{X}}}$ is an orthogonal projection matrix for $\widetilde{\mathbf{X}}$.[27] We show in Appendix C that $\widehat{\theta}_{IV}$ is an unbiased estimator of θ if our vector of instruments, $\mathbf{Z}$, satisfies the following moment condition:

$$\mathbb{E}\left[c_v^U \;\middle|\; \mathbf{Z}, \widetilde{\mathbf{X}}, \log \text{density} \right] = 0\,. \tag{5}$$

Crucially, the instruments need to be uncorrelated with omitted community-level factors that influence overall social capital in the community.

Soil Characteristics as Instruments for Density. We propose that within urban areas, deep soil characteristics satisfy the moment condition in equation (5). To measure soil attributes, we use data from SoilGrids to capture various features of the different soils predominant in Indonesian communities, including (1) bulk density; (2) water content; (3) sand content; (4) clay content; (5) texture; and (6) soil taxonomy information. Stable, fertile soils historically attracted greater numbers of people to settle in specific areas, affecting both traditional societies and also colonial investments (Dell and Olken, 2019). We show below that within metropolitan areas, certain soil characteristics also have a strong first stage

[25] More subtly, group averages will also absorb part of the unobserved group quality component that is both orthogonal to observed group characteristics and correlated with amenities that families consider when choosing where to live.

[26] This insight was actually discussed by Altonji and Mansfield (2018). From p. 2094, with emphasis added: "... [T]he fact that controlling for the group averages eliminates bias from sorting implies that the causal effects ($\mathbf{\Gamma}$) of *particular school inputs or policies* (in $\mathbf{Z}_s$) can be *point identified* in situations where bias from omitted neighborhood/school characteristics in z_s^U is not a problem or can be addressed through a *complementary instrumental variables* scheme."

[27] This matrix is given by: $\mathbf{M}_{\widetilde{\mathbf{X}}} = \mathbf{I} - \widetilde{\mathbf{X}}\left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'$.

relationship with population density today. Similar geologic instruments for density have also been used in prior work (e.g Hoxby, 2000; Black et al., 2002; Rosenthal and Strange, 2008; Combes et al., 2010).

Despite the first stage relationship, there are several concerns with using soil characteristics as an instrument for density. One issue could be that soil characteristics recorded in cities today may reflect human activity, so these instruments could introduce reverse causality or simultaneity concerns. As discussed in Section 3, we only use soil attributes measured at a depth of 60 cm or more, helping to ensure that they are unaffected by human activity. We also do not consider certain measures that are easily changed by human activity, such as organic carbon content and soil acidity (pH).

A second, larger concern is the exclusion restriction, namely that within cities, soil attributes need to only affect social capital outcomes today through their effect on density. Even though soil mineralogy and the parent materials of soils were determined millions of years ago, they may still be relevant drivers of local wealth within cities, particularly if parts of urban areas contain agricultural employment. In the analysis that follows, we take care to examine this and several other potential threats to the exclusion restriction.

Sorting Controls and the Interpretation of θ. Note that when we introduce sorting controls in equation (3), we are estimating the effect of density on social capital conditional on the spatial distribution of the population. If social capital accumulates dynamically, historical sorting patterns that affect density today could also affect social capital outcomes. When we include the $\mathbf{X}_v$ sorting controls, this effectively removes this variation, and what we are left with is the contemporaneous effect of density on social capital, conditional on the sorting patterns of people with different demographic characteristics.

We think that θ is useful for studying how policies that shape the density of a community may impact social capital. For example, in the U.S. and other developed countries, policymakers use many different regulations to control the levels of density that emerge in a community. Such policies include: (1) binding limits on new construction; (2) open space dedications; (3) growth controls; (4) environmental regulations; (5) septic system regulations; (6) subdivision requirements; and (8) historic preservation (Glaeser and Ward, 2009). These policies alter the levels of density that are allowed to emerge in a community, but they do not directly control who gets to live where. Our estimates capture the extent to which density impacts different aspects of social capital conditional on the spatial distribution of the population, which is more relevant for policy.

6 Results

First Stage. We begin by using post-double-selection lasso techniques to select soil characteristics that are the best predictors of log density in 2010, following Belloni et al. (2012). Table 2 reports parameter estimates from the following regression equation:

$$\text{log density}_v = \alpha_c + \mathbf{z}_v'\beta + \mathbf{C}_{2v}\theta + \varepsilon_v \, , \tag{6}$$

where v indexes communities, c indexes urban areas, α_c denotes a city-specific intercept, $\mathbf{z}_v$ denotes the selected vector of soil characteristics, $\mathbf{C}_{2v}$ denotes an additional vector of exogenous community-level characteristics (including ruggedness, elevation, and distance to the nearest coast and river), and ε_v is an error term.

Out of 67 candidate soil characteristics instruments for density, all measured at a depth of 60 cm or more, the post-double-selection lasso procedure selected only 6 instruments, and column 1 of Table 2 reports their first-stage coefficients. The overall F-statistic for the regression is large (91.2), and within urban areas, the regression explains roughly 36 percent of the variation in density. We find that within urban areas and conditional on $\mathbf{z}_v$ controls, population density in 2010 was positively related to the bulk density of the soils' parent material. This seems reasonable, given that more compact soils provide favorable land for construction. We also find that sand content is negatively related to density. Although sandy soils may be favorable for construction, they are also difficult for growing crops and likely reduced historical agricultural productivity.[28]

In addition, four soil types were also significant predictors of population density. Haplustolls (from the order Mollisols) are grassland soils that are often used for growing grains and feed crops (USDA, 2015). These areas were probably locations that were initially favorable for rice production and influenced historical settlement patterns, and as expected, the density coefficient on Haplustolls is positive. Both Haplorthox (from the order Oxisols) and Tropodults (from the order Ultisols) are soils predominant in tropical forests, and we find that within urban areas, these soil types reduce density, as may be expected (USDA, 1975). Such areas were more difficult for growing crops historically, and as urban areas have sprawled and land has been cleared, they are home to lower-density urban development. Dystropepts (from the order Inceptisols) are often found in mountainous areas and on steep slopes (USDA, 2015), but even conditional on elevation and ruggedness, we find that Dystropepts reduces density. In summary, the selected soil characteristics within cities are either associated with attracting historical development through favorable agricultural production or are associated with the ease and facility of clearing land and constructing buildings.

In Column 2, we add controls for sorting on observables and unobservables, $\mathbf{X}_v$, to equation (6). The coefficients on the selected soil characteristics retain their signs and statistical significance. Within urban areas, the combination of soil characteristics, exogenous physical characteristics, and sorting controls explain 70 percent of the variation in density. Overall, the results from Table 2 suggest that conditional on city fixed effects, the selected soil characteristics have a strong first stage relationship with community-level population density, the key dependent variable in our analysis.

Baseline Results. To estimate the impact of density on social capital outcomes, we run linear instrumental variable regressions of the following form:

$$y_{vi} = \alpha_c + \mathbf{X}_i\beta + \theta \log \text{density}_v + \mathbf{X}_v\mathbf{\Gamma}_1 + \mathbf{C}_{2v}\mathbf{\Gamma}_2 + \varepsilon_{vi}, \tag{7}$$

where α_c denotes city fixed effects, $\mathbf{X}_i$ is a vector of individual-level observables, $\mathbf{X}_v$ is a vector of 38 group averages of individual characteristics of people living in community v (described above), $\mathbf{C}_{2v}$ are community characteristic controls that are not mechanically related to sorting, and ε_{vi} is an error term. To illustrate our empirical strategy, Table 3 shows the results for a single outcome variable, namely levels

[28] Although elevation, ruggedness, and distance to the coast and rivers are also correlated with density as one might expect, we do not use them as instruments because they could directly impact social capital today. For example, living in a more rugged community could make it more difficult to communicate with one's neighbors. Similarly, living in a higher elevation community might increase the costs of community participation.

of the 4-point trust in neighbors index. Standard errors are robust and clustered at the sub-district level.

In Panel A of Table 3, we report estimates of θ from separate specifications, where we omit controls for sorting and set $\mathbf{\Gamma}_1 = 0$ as a baseline. In column 1, our OLS specification finds that increasing log density by 1 (or increasing density by 2.7 people per km) reduces the trust in neighbors index by 0.026 points. Although highly significant, this is a moderate effect size, equivalent to roughly 5 percent of a standard deviation in the index.

Column 2 reports the relationship between density and trust in neighbors estimated from IV-Lasso specifications. Overall, the estimate of θ remains highly significant but increases in absolute terms to -0.056.[29] In column 2, the Kleibergen and Paap (2006) Wald Rank F-Stat, a generalization of the first-stage F-statistic for multiple instrumental variables, is large at over 82. Both the Kleibergen-Paap LM test and the Anderson-Rubin (AR) test strongly reject the null of weak instruments for the endogenous density variable. Finally, the Sargan-Hansen J-test statistic for overidentifying restrictions is small, and we cannot reject the null that the soil characteristics instruments are correctly excluded from the estimation equation. Overall, the results in Table 3 point to a well-specified IV model. The fact that our IV estimates are more negative than the OLS estimates suggests that the least squares estimates are positively biased. This could be due to omitted, community-specific factors that both increase trust and attract greater numbers of people, leading to positive simultaneity bias.

In Panel B, we report results of the full model, where we include individual-specific variables, $\mathbf{X}_i$, community-specific characteristics, $\mathbf{C}_{2v}$, and averages of 38 different individual-level variables at the community level, $\mathbf{X}_v$, to control for sorting. In column 2, the estimated effect of density on trust in neighbors increases slightly and remains significant at conventional significance levels. Although the Kleibergen-Paap Wald Rank F-Stat falls in the Panel B specifications, the Kleibergen-Paap LM tests still reject the null of weak instruments of the endogenous density variable. Moreover, the Sargan-Hansen J-test statistic falls further, suggesting that the IV model is still well specified, even after introducing controls for sorting.

At the bottom of the table, we report p-values of F-tests on the significance of $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$. These tests reject the null that they are jointly equal to zero, which suggests that controls for sorting and controls for community-level covariates matter for predicting outcomes. We also run an F-test that compares θ_A, the estimate of θ from Panel A to θ_B, the estimate of θ from Panel B. In the IV specification, we cannot reject the hypothesis that these two estimates are equal. Overall, the results from Table 3 suggest that greater density moderately reduces trust in neighbors, and that this effect is causal and robust to controls for sorting on observables and unobservables.[30]

Mean Effects Results. To estimate the effect of density on multiple different dimensions of social capital, we create summary impact measures using a mean effect analysis, following Kling et al. (2007). To estimate mean effects, we form groups of related outcomes, where a single outcome for individual i in community v is given by $y_{iv}(k)$, and $k = 1, ..., K$ indexes outcomes. We modify the signs of each variable

[29] We report the individual-level first stage results in Appendix Table A.3. We also suppress estimates of β and $\mathbf{\Gamma}_2$ from Panel A, Table 3, but Appendix Table A.4 reports these coefficients. Trust in neighbors falls with education and has an inverse U relationship with age.

[30] In Appendix Table A.5, we report the full set of estimates of $\mathbf{\Gamma}_1$ from the specification in Panel C, Table 3 (suppressing estimates of β and $\mathbf{\Gamma}_2$). A greater share of recent migrants reduces trust in neighbors, while an increased share of "ever migrants" increases trust in neighbors. The sizes of shares of certain ethnicities are also related to trust in neighbors.

in the group so that increases denote greater social capital (e.g., improved trust, increased community or social insurance participation, or greater intergroup tolerance). Next, we simultaneously estimate (7) for all K outcomes, using a SUR system and a stacked vector of the standardized $y(k)$'s as the dependent variable. The mean effect size we report is simply an estimate of the weighted average effect of density on this group of outcomes, where each separate effect is weighted by the outcome's standard deviation. Formally, this is given by:

$$\tau = \frac{1}{K}\sum_{k}\frac{\theta_k}{\sigma_k}\,, \tag{8}$$

where K is the total number of outcomes in the grouping, θ_k is the effect of density on outcome k (measured from a regression akin to equation (7)), and σ_k is the standard deviation of outcome k.[31]

In Table 4, we report estimates of τ for different groups of social capital outcomes. Each cell in this table reports a different mean effect size. Different sets of columns are reserved for the four different outcome groupings (as described in Table 1), while different rows are used to include or exclude the $\mathbf{X}_v$ sorting controls. From the results in columns 1-2, we see that greater density reduces trust in neighbors, and this effect is robust to controls for sorting (row 2).

One way of interpreting this effect size is to consider how trust in neighbors changes when moving from an average neighborhood in the suburbs to an average neighborhood near the CBD. On average, log density near the CBD for cities in our sample is approximately 8.8, while around 20 km away, log density declines to 6.7. So, we can multiply $\widehat{\tau}$ by 2.1 to find that moving from an average suburban neighborhood to one near the CBD reduces average trust in neighbors by roughly 0.25 standard deviations (σ). This moderate effect size is substantially larger than, for example, the impact of an additional year of schooling on average trust in neighbors (-0.002 σ).

In columns 3 and 4, we also find that greater density is associated with reduced community participation. This effect is robust to including $\mathbf{X}_v$, suggesting that sorting and simultaneity do not confound the relationship. Increasing density by moving 20 km from a suburban community to a community by the CBD causes a 0.15 σ reduction in community participation, a moderate effect size. This effect size is also much larger than the impact of an additional year of schooling on community participation (0.008 σ). This effect could have important policy implications, and efforts to increase participation in higher-density communities, perhaps through greater outreach or community organizing, may reverse these trends. The findings of columns 2 and 4 echo those of Brueckner and Largey (2008) on the negative effects of density on social interactions in U.S. cities.

Based on the F-test results reported at the bottom of the table, for trust and community participation, we can reject the joint hypothesis that the coefficients on the $\mathbf{X}_v$ controls are equal to zero, so these sorting controls clearly matter for explaining variation in outcomes. However, in all IV specifications, we cannot reject the hypothesis that estimates of τ from row 1 are different from row 2. This suggests that the relationship between density and trust in neighbors is causal and robust to both controls for sorting and simultaneity.

[31] To calculate the mean effect size τ and compute standard errors, we follow the supplementary appendix of Kling et al. (2007) and use a seemingly unrelated regressions (SUR) system to estimate the effect of population density on outcomes. Standard errors are obtained from the variance-covariance matrix of the SUR system. This approach allows us to estimate a single mean effect across all individuals in the 2012 *Susenas*, even when missing values are reported for certain responses. More details can be found in Appendix C.5.

In columns 5 and 6, we investigate the relationship between density and participation in social insurance in the community. Social insurance is an important way that lower income households share risks (Townsend, 1994), and the strength of social insurance ties has also been shown to reduce migration to cities in other contexts (Munshi and Rosenzweig, 2016). However, we do not find a robust or statistically significant relationship between density and social insurance, suggesting that this important dimension of social capital may not be affected by the overall density environment.

Finally, in columns 7 and 8, we find that greater density is positively associated with intergroup tolerance, but this effect is not robust to controls for sorting (row 2). This finding suggests that individuals who dislike other ethnic groups may be sorting into less dense and more homogeneous areas, but density itself does not seem to affect intergroup tolerance. Such a finding is potentially interesting given the increase in sorting and political polarization in many nascent democracies, such as Indonesia.[32]

6.1 Probing Instrument Validity

Our identification strategy relies on the assumption that the soil characteristics IVs we selected affect social capital outcomes today only through their effects on density. In this subsection, we take care to rule out several plausible channels through which the IVs may independently affect social capital outcomes and violate the exclusion restriction.

Agricultural Productivity and Employment. If favorable soils were important for agricultural productivity in cities today, they could influence social capital by affecting income, wealth, or social relationships. This is clearly a concern, since roughly 23 percent of the households in our sample are employed in agriculture.[33] In Appendix Table A.8, we show that conditional on urban area fixed effects, community characteristics, and sorting controls, the selected soil characteristics that we use to predict population density are not individually significant in predicting rice productivity, food crop productivity, cash crop productivity, or total agricultural productivity.[34] Furthermore, a lasso procedure on the full set of soil characteristics fails to identify any soil characteristics that can predict rice productivity, cash productivity, or total agricultural productivity in urban areas (Appendix Table A.9).

Nonetheless, we may still be concerned about the role of agricultural employment in explaining our results. In Appendix Table A.10, we show that our mean effects results are robust to excluding agricultural households from the sample and to dropping communities with a significant share of agricultural employment. Furthermore, a sensitivity analysis presented in Appendix Table A.11 shows that the estimated effects of density are robust to dropping peripheral communities that are far from the city center. Overall, these results provide reassurance that the IVs we use did not affect social capital by impacting contemporaneous agricultural activities.

[32] The individual linear-index outcome results for trust and community participation, upon which estimates of τ in Table 4, Panel A and Panel B are based, can be found in Appendix Tables A.6. The same results for social insurance and inter-ethnic tolerance can be found in Appendix Table A.7.

[33] Appendix Figure A.5 shows that agricultural employment rises as distance to the central business district increases and density falls, as one might expect. This trend is apparent regardless of whether agricultural employment is measured with 2010 census data (Panel A) or with the 2012 Susenas data (Panel B).

[34] To measure agricultural productivity, we used yields data from the 2002 Indonesian Village Potential Survey (or *Podes*) and national crop prices from FAO/PriceStat data. The crop categories include: (1) rice; (2) secondary food crops, known collectively as *palawija*, which include maize, cassava, groundnuts, sweet potato, and soybeans; (3) cash crops, the most important of which are palm oil, rubber, cocoa, and coffee; and (4) total agricultural production, which includes all crops.

Persistent Impacts of Historical Infrastructure and Culture. Even though our soil characteristics IVs do not predict agricultural productivity in urban areas, they might be correlated with other omitted location characteristics that drive both historical populations and explain social capital outcomes today. Although we control for several aspects of favorable geography (e.g., ruggedness, elevation, distance to the coast and rivers), other omitted place-based characteristics could be correlated with density and may have also contributed historically to the development of social and physical infrastructure. Similarly, the IVs may also be correlated with characteristics that drive the formation of cultural and social norms we observe today. For example, Alesina et al. (2013) show that soil characteristics historically affected the use of the plough, which had persistent impacts on gender norms.

In Appendix Table A.14, we show that our results are robust to controlling for these potential channels. Column 3 shows that our mean effects results are robust to adding controls for many different types of physical and social infrastructure that were built or established before 1983.[35] Column 4 shows that our mean effect estimates are also robust to controlling for the share of households in the community in 2000 who belong to an ethnicity that practices different cultural norms as documented in the *Ethnographic Atlas*.[36]

Figure 5 summarizes our results so far. For each group of social capital outcomes, we plot the point estimate and confidence interval for the IV-Lasso specifications with sorting controls (akin to Table 4, row 2). We plot confidence intervals for: (i) the full sample baseline; (ii) estimates dropping agricultural households; (iii) estimates dropping communities with the share of agricultural households exceeding 20 percent; (iv) estimates controlling for historical infrastructure; and (v) estimates controlling for culture and social norms. Estimates of τ and their confidence intervals are remarkably similar across these specifications, which suggests that our estimates are unlikely to be confounded by contemporaneous agriculture, historical infrastructure, historical social organizations, or culture.

A Placebo Exercise. Finally, as more general evidence in favor of the exclusion restriction, we performed a placebo exercise and estimated the effect of our soil characteristics IVs on social capital outcomes in communities where those IVs do not predict density. The intuition behind this exercise is that in a subsample where we have no first stage relationship, there should also be no reduced-form relationship if the exclusion restriction is satisfied (Altonji et al., 2005; van Kippersluis and Rietveld, 2018).

We implemented this placebo exercise in two steps. First, we focused on rural areas where the relationship between our IVs and population density is weak. We use the UN Statistical Commission's definition of rural areas as locations where population density is less than 300 inhabitants per square km

[35] We constructed variables from the 1983 Village Potential Survey (*Podes*) to measure several aspects of infrastructure investments and community organizations. These historical controls include: (1) education facilities (the number of kindergarten, primary, junior secondary, and senior secondary schools); (2) medical facilities (the number of hospitals, the number of community health clinics, or *Puskesmas*, and the number of community-based preventative and promotive care facilities or *Posyandu*); (3) the number of places of worships (counts of the number of mosques, surau, churches, pura, and vihara); (4) irrigation infrastructure (the share of wet/paddy rice fields that use man-made irrigation); (5) utilities (share of households covered by the national electricity grid); (6) the share of communities with various agricultural and social organizations; and (7) distance to major roads.

[36] The culture and social norms controls we use include the share of households in the community who: (1) make a "bride price" payment at the time of marriage; (2) are from matrilocal societies; (3) are from patrilocal societies; (4) traditionally practiced male-led agriculture; (5) traditionally practiced female-lead agriculture; (6) practiced slavery historically; and (7) traditionally practiced polygamy. To control for differences in culture, we merge these *Ethnographic Atlas* variables, which vary at the ethnicity level, to the 2000 census, recorded 12 years before our social capital outcomes are measured.

(UN Statistical Commission, 2020).[37] Next, we estimated the reduced-form relationship between our IVs and social capital outcomes in this subsample of communities.

Appendix Table A.15 shows that in communities where the IVs do not predict density, there is no significant reduced-form relationship between the IVs and nearly every social capital outcome we study. At the bottom of the table, we show that the first-stage relationships between the IVs and population density in these communities are weak, with Kleibergen-Paap Wald F-stats of around 4. The table also reports p-values for tests of the joint significance of the soil characteristics for each dependent variable. Out of 12 social capital outcomes, the selected soil characteristics are significantly related to only a single community participation variable. We take this, along with the other results in this section, as evidence strongly in favor of the exclusion restriction.

6.2 Additional Robustness Checks

Specification Checks. In Appendix B, we show that our main results are robust to a number of different specification checks. We first coarsen the multivalued dependent variables into binary indicators and estimate the effects of density with linear probability models and IV probit specifications. Next, we use ordered probit models with instruments, adapting the control function procedure of Chesher and Rosen (2019). Our results on the impact of density on individual outcomes are robust to these different specification choices.

Additional Community-Level Controls. Next, we explore whether our estimated effects are really due to density, or owe instead to other amenities that are influenced by density. To do so, we use the 2011 Podes data to add additional community-level variables to the $\mathbf{C}_{2v}$ vector. We construct several proxies for local amenities that may influence social interactions, including the community's distance to formal markets, if any restaurants exist, distance to schools, if there are any mobile phone or TV signals, the type of main water sources, if there are local community empowerment programs, the number of houses of worships, distance to medical facilities, and distance to maternal health facilities. Comparing rows 2 and 3 across the panels of Appendix Table A.16, we find that these additional controls do not affect our main estimates from Table 4.

IFLS Results. Next, we estimate the impact of density on social capital using a different dataset: the Indonesia Family Life Survey (IFLS). The IFLS is a national longitudinal survey that is representative of 83 percent of Indonesia's population. It tracks more than 30,000 individuals in 5 waves over a 19-year period. The IFLS provides a useful check against our main *Susenas* results for several reasons. First, the IFLS contains different social capital measures and a completely different sampling strategy from the *Susenas*, so it would be reassuring if we found similar effects in the cross-section. Second, we can exploit individual-level panel data from the IFLS to address sorting in a completely different way from what we do with the cross-sectional *Susenas* results.

[37] Although BPS has definitions for rural and urban areas in Indonesia, many communities that BPS classifies as rural are actually quite densely populated. Of the 6,751 communities in the 2012 *Susenas*, 4,208 communities were either classified as rural by the UN density threshold or by BPS. A total of 1,626 (38.6%) were categorized as rural by BPS but had density larger than 300 km. Only 201 communities (4.7%) were classified as rural based on population density but urban based on BPS definitions.

We use data from waves 4 (2007) and 5 (2014) that contain the complete social capital module.[38] We grouped the variables from the social capital module into the four categories used in Table 4.[39] Next, using restricted access data, we linked IFLS communities to communities from the 2010 census, so that we could obtain density measures and controls for sorting. Finally, we estimated the mean effects of density on social capital outcomes, analogous to Table 4.

Table 5 reports IFLS results on the single cross-section from 2015. In columns 1 to 4, we find very similar estimates of the effect of density on trust in neighbors and community participation compared to what we report in Table 4. In columns 5 and 6, we find that density is negatively related to participation in social insurance, and these results are statistically significant, unlike those reported in Table 4. We suspect that some of the differences here could be due to differences in how the single question on social insurance in the IFLS was worded, compared to the 3 different questions asked in the 2012 *Susenas* survey. Finally, we find larger estimates of the impact of density on intergroup tolerance, but again they are not robust to controls for sorting. Overall, we view these results from IFLS 5 as broadly consistent with the qualitative patterns found in our main *Susenas* results.[40]

Next, we use the longitudinal nature of the IFLS to address sorting in a different way from our cross-sectional results, following the two-step estimation approach described by Combes et al. (2008). In the first step, we estimate local, time-varying effects of social capital after conditioning out individual fixed effects and time-varying individual-level observables. This step effectively purges the social capital outcomes of any bias from sorting. We then average the residuals from this regression over community years, and estimate a cross-sectional regression of the average social capital measures on our density measure in 2010.[41]

For a single outcome, the first step involves estimating the following regression equation:

$$y_{ivt} = \mathbf{x}_{it}'\beta + \alpha_i + \alpha_{vt} + \varepsilon_{it}\,, \tag{9}$$

where y_{ivt} is the social capital outcome for individual i in community v in year t, $\mathbf{x}_{it}$ is a vector of time-varying controls for individual i (capturing age, changes in education, and changes in marital status), α_i is an individual fixed effect, α_{vt} is a community-year intercept, and ε_{it} is an error term. The object of interest in this regression is α_{vt}, which is the social capital index for each community and year, after conditioning out individual fixed effects and time-varying individual-level observables. Because we work with a large number of related outcomes, we use a mean effects approach and estimate equation (9) with a stacked SUR system, where we impose the restriction that the α_{vt} terms are common across equations.

In the second step, we form a community-level average of the α_{vt}'s across years, $\alpha_v \equiv \frac{1}{T}\sum_{t=1}^{T}\alpha_{vt}$,

[38] Although the IFLS has 5 waves to date, questions on trust and intergroup tolerance were only asked in waves 4 and 5. While community participation questions were also asked in wave 3, we only use waves 4 and 5 to ensure consistency across outcomes.

[39] The variable names, groupings, and summary statistics for the IFLS 5 (IFLS 4) can be found in Appendix Table A.17 (Appendix Table A.18). Both sets of social capital questions were worded identically between the two waves.

[40] Appendix Table A.19 presents cross-sectional mean effects results for IFLS 4, using density in 2010 as the dependent variable. These results have similar magnitudes, but they are somewhat less robust to the sorting controls.

[41] See Appendix C.6 for more precise details on how this approach was implemented.

and we use this as the dependent variable in a cross-sectional regression:

$$\alpha_v = \mathbf{C}_{2v}\beta_2 + \theta \log \text{density}_v + \Delta\varepsilon_i \,, \tag{10}$$

where we instrument log density$_v$ with the soil characteristics instruments, and $\mathbf{C}_{2v}$ is defined as above. We restrict the sample to contain only the original 182 IFLS communities in urban areas in Indonesia.[42]

Table 6 reports our two-step estimates of θ from the IFLS panel data. Although our estimates are generally not significant, they have similar signs and magnitudes as those reported in Table 4. A major difference between these two specifications is that the IFLS panel results are based on a much smaller sample size, increasing confidence intervals and reducing the power of the instruments. Nevertheless, we take the evidence in Table 6 as broadly consistent with our main findings.[43]

6.3 Heterogeneity and Mechanisms

Although Putnam (2000) and others have argued that low density development in the periphery of cities could reduce social capital due to greater commuting times and increased opportunity costs of social interactions, we find the opposite result (Glaeser et al., 2002; Glaeser and Gottlieb, 2006). One set of theories that could explain these results is that crowding in higher densities is exhausting (Simmel, 1903; Wirth, 1938), increasing mistrust as social interactions become more anonymous (Brueckner and Largey, 2008; Habyarimana et al., 2007). Alternatively, high density may be associated with greater criminal activity, making people more suspicious and reluctant to participate in their communities.

To explore these theories, in Table 7, we investigate how different individual and city-level characteristics mediate the relationship between density and social capital. To estimate heterogeneous effects of density on a single outcome, we specify the following regression equation,

$$y_{vi} = \mathbf{X}_i\beta + \theta \log \text{density}_v + \theta_0\, m_i + \theta_1 \left(\log \text{density}_v \times m_i\right) + \mathbf{X}_v\mathbf{\Gamma}_1 + \mathbf{C}_{2v}\mathbf{\Gamma}_2 + \varepsilon_{vi}, \tag{11}$$

where m_i is an individual-level (or city-level) binary variable, and the other terms in the equation are defined as above.[44] In Table 7, we estimate stacked SUR systems of equations like (11) for different groups of outcomes, and report estimates of the mean effects of density and density's interaction with m_i.

Panel A of Table 7 reports our baseline estimates of the mean effects of density from Table 4. In Panel B, we examine whether the effects of density vary for individuals with different levels of education,

[42] Although there were 2,330 communities observed in IFLS 3 and 3,343 communities observed in IFLS 4, many of these communities correspond to only a handful of observations, as those communities are where individuals from the original IFLS communities moved and formed new households. Working with those communities makes it difficult to reliably estimate α_{vt} separately from individual-level fixed effects. There are a total of 312 original communities in the IFLS but only 182 were located in our urban area sample. For the 21,436 individual observations in this sample, roughly 9 percent were movers either into these 182 villages or away from them. A total of 1,097 individuals (5.12% of the sample) moved into the 182 villages, while 931 (4.34% of the sample) moved out. A histogram of the distribution of the number of observations used to estimate the α_{vt} terms from equation (9) in the first step of the procedure can be found in Appendix Figure A.6; the median number of observations is 114, but there is considerable variation across villages and years.

[43] In the first stage of Table 6, we use a SUR approach, but we also try a single-index approach in the first step, forming a single average of the dependent variables in each group as the regressor. Estimates of θ from the second step using this single-index approach can be found in Appendix Table A.20. These results are broadly similar to those presented in Table 6.

[44] The set of instruments we use is augmented by interactions between our instruments and those indicators. See Appendix C.7 for more details. The interaction variables and their sources are also described in Appendix Table A.21.

income, or ownership of a private transportation mode. The first set of rows shows that higher education amplifies the negative impacts of density on community participation and social insurance. In the second set of rows, we find similar, albeit much noisier, results with regards to higher income. These results are consistent with the hypothesis that density's effects on crowding are stronger for wealthier and more educated people. Moreover, if density increases the marginal time costs of community participation, for example because of longer community meetings in dense areas, individuals with higher opportunity costs may be the first to cease attendance.

On the other hand, we do find suggestive evidence for the Putnam (2000) hypothesis when we look at intergroup tolerance. Column 4 of Panel B suggests that density is somewhat positively associated with interethnic tolerance among those with higher opportunity costs. However, this effect is only precisely estimated for higher income earners.

In the third set of rows of Panel B, we find that the effects of density on trust in neighbors, community participation, and social insurance are more negative for people who take private transport modes (cars and motorcycles), relative to the reference group of public transport and non-motorized transit. This provides an interesting test of the hypothesis that increased commuting costs (in sprawling communities) reduced investments in social capital (Putnam, 2000). On the one hand, all else the same, having private transport should reduce commuting time. However, by allowing more convenient travel outside of one's locality, it also increases the opportunity costs of investing in (local) social capital. Our results suggest that the latter effect dominates the former.

In Panel C, we investigate the extent to which city-level factors may enhance or alter the impact of density on social capital. The first set of rows looks for differences in the effects of density across cities that have longer commute distances, measured from labor force survey data (*Sakernas*). Despite increasing the opportunity costs of community participation, we do not find a less negative density impact in cities with longer commutes. Instead, we find suggestive evidence that cities with greater commuting distances have more negative effects of density on social capital, calling into question the opportunity cost story of Putnam (2000).

In the final two sets of rows, we examine the role of crime in mediating the results on density and social capital outcomes. We use the 2011 *Podes* data to construct measures of the probability of encountering violent and property crimes in each city. We then interact density with indicators for whether or not a city has an above-median exposure to property or violent crime. We generally find that the effects of density are amplified in higher crime cities. Exposure to crime in such cities might reduce the benefits and increase the costs of community participation, encouraging them to sort into enclaves in lower-density areas, where they invest more heavily in their communities and have a lower tolerance for other ethnic and religious groups.

7 Conclusion

This paper presents causal estimates of the effect of urban sprawl on different aspects of social capital in Indonesian cities. Researchers who estimate these relationships must address two fundamental identification problems: (1) simultaneity, in which omitted place-specific variables drive both density and social capital, and (2) sorting, where individuals with particular tastes for contributing to social capital systematically sort into places with different levels of density. Using high-quality, spatially disaggregated data, we confront the first identification challenge by instrumenting for density within urban areas using soil characteristics. We address the second challenge using controls for sorting on observables and unobservables, extending an approach by Altonji and Mansfield (2018).

Our major finding is that in Indonesian cities, increases in density lead to lower levels of trust in neighbors and reduced community participation, echoing the results of Brueckner and Largey (2008) for U.S. cities. These results are robust to multiple threats to the exclusion restriction, different datasets, and to two different approaches to addressing sorting. We also find that increased density leads to greater levels of intergroup tolerance, but these effects are not robust to sorting controls. Moreover, our heterogeneity analysis shows that the effect of density is amplied for higher income and more educated individuals. The negative relationship between density and social capital is also moderately more pronounced in higher crime cities.

As emphasized by Brueckner and Largey (2008), social planners may want to use growth controls to curb sprawl if it leads to undesirable social externalities. For instance, if sprawl were to causally reduce interethnic tolerance, this could provide a rationale for policy intervention. However, lower density also has positive social externalities because it increases trust in neighbors and community participation. Because we understand very little about the tradeoffs between improving within-community cohesion at the expense of fostering intergroup relationships, policymakers should proceed with caution. These social capital impacts also need to be compared to the costs of other aspects of urban sprawl, especially energy use and the carbon intensity of living.

More research is needed to understand whether these patterns of sprawl and social capital are common to LMICs. Indonesia’s unique history and relatively weak interethnic conflict make it an interesting case study, but the impact of sprawl on social capital could be very different in countries with a legacy of violent ethnic conflict, greater religious tensions, or a recent experience of civil war. Other aspects of urban sprawl, particularly energy use and the carbon intensity of living, are also first order in trying to quantify the costs and benefits of sprawl in LMIC cities.

References

AGRAWAL, M., J. G. ALTONJI, AND R. K. MANSFIELD (2019): "Quantifying family, school, and location effects in the presence of complementarities and sorting," *Journal of labor economics*, 37, S11–S83.

ALESINA, A., P. GIULIANO, AND N. NUNN (2013): "On the origins of gender roles: Women and the plough," *The quarterly journal of economics*, 128, 469–530.

ALESINA, A. AND E. LA FERRARA (2000): "Participation in Heterogeneous Communities," *Quarterly Journal of Economics*, 115, 847–904.

——— (2002): "Who trusts others?" *Journal of Public Economics*, 85, 207–234.

ALESINA, A. AND E. ZHURAVSKAYA (2011): "Segregation and the Quality of Government in a Cross Section of Countries," *American Economic Review*, 101, 1872–1911.

ALTONJI, J. G., T. E. ELDER, AND C. R. TABER (2005): "An evaluation of instrumental variable strategies for estimating the effects of catholic schooling," *Journal of Human resources*, 40, 791–821.

ALTONJI, J. G. AND R. K. MANSFIELD (2018): "Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects," *American Economic Review*, 108, 2902–46.

ANGEL, S., J. PARENT, D. L. CIVCO, A. BLEI, AND D. POTERE (2011): "The dimensions of global urban expansion: Estimates and projections for all countries, 2000–2050," *Progress in Planning*, 75, 53–107.

ASHRAF, N., N. BAU, N. NUNN, AND A. VOENA (2020): "Bride Price and Female Education," *Journal of Political Economy*, 128, 591–641.

ASIAN DEVELOPMENT BANK (ADB) (2012): "Green Urbanization in Asia: Key Indicators for Asia and the Pacific 2012," Technical report, Asian Development Bank.

BARRON, P., K. KAISER, AND M. PRADHAN (2009): "Understanding Variations in Local Conflict: Evidence and Implications from Indonesia," *World Development*, 37, 698–713.

BAZZI, S., A. GADUH, A. D. ROTHENBERG, AND M. WONG (2019): "Unity in Diversity? How Intergroup Contact Can Foster Nation Building," *American Economic Review*, 109, 3978–4025.

BEARD, V. A. AND A. DASGUPTA (2006): "Collective Action and Community-driven Development in Rural and Urban Indonesia," *Urban Studies*, 43, 1451–1468.

BEBBINGTON, A., L. DHARMAWAN, E. FAHMI, AND S. GUGGENHEIM (2006): "Local Capacity, Village Governance, and the Political Economy of Rural Development in Indonesia," *World Development*, 34, 1958–1976.

BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

BERTRAND, J. (2004): *Nationalism and ethnic conflict in Indonesia*, Cambridge, UK; New York: Cambridge University Press.

BESSER, L. M., M. MARCUS, AND H. FRUMKIN (2008): "Commute time and social capital in the US," *American Journal of Preventive Medicine*, 34, 207–211.

BISHOP, B. (2009): *The big sort: Why the clustering of like-minded America is tearing us apart*, Houghton Mifflin Harcourt.

BLACK, D., K. DANIEL, AND S. SANDERS (2002): "The impact of economic conditions on participation in disability programs: Evidence from the coal boom and bust," *American Economic Review*, 92, 27–50.

BOSKER, M., J. PARK, AND M. ROBERTS (2021): "Definition matters. Metropolitan areas and agglomeration economies in a large-developing country," *Journal of Urban Economics*, 125, 103275.

BOWEN, J. R. (1986): "On the political construction of tradition: Gotong Royong in Indonesia," *The Journal of Asian Studies*, 45, 545–561, publisher: Cambridge University Press.

BRUECKNER, J. K. AND A. G. LARGEY (2008): "Social interaction and urban sprawl," *Journal of Urban Economics*, 64, 18–34.

BRUECKNER, J. K. AND K. S. SRIDHAR (2012): "Measuring welfare gains from relaxation of land-use restrictions: The case of India's building-height limits," *Regional Science and Urban Economics*, 42, 1061–1067.

BURCHFIELD, M., H. G. OVERMAN, D. PUGA, AND M. A. TURNER (2006): "Causes of sprawl: A portrait from space," *The Quarterly Journal of Economics*, 121, 587–633.

CALDEIRA, T. P. R. (2001): *City of Walls: Crime, Segregation, and Citizenship in São Paulo*, University of California Press.

CAROTHERS, T. AND A. O'DONOHUE (2019): *Democracies divided: The global challenge of political polarization*, Brookings Institution Press.

CHESHER, A. AND Z. ROSEN, ADAM M.AND SIDDIQUE (2019): "Estimating endogenous effects on ordinal outcomes," CEMMAP working paper.

CHOMITZ, K. M., P. BUYS, AND T. S. THOMAS (2005): "Quantifying the Rural-Urban Gradient in Latin America and the Caribbean," Policy Research Working Paper 3634, Development Research Group, Infrastructure and Environment Team, World Bank, Washington, DC.

CIVELLI, A. AND A. GADUH (2018): "Determinants of Urban Sprawl: Evidence from Indonesia," .

COMBES, P.-P., G. DURANTON, AND L. GOBILLON (2008): "Spatial wage disparities: Sorting matters!" *Journal of urban economics*, 63, 723–742.

COMBES, P.-P., G. DURANTON, L. GOBILLON, AND S. ROUX (2010): "Estimating Agglomeration Economies with History, Geology, and Worker Effects," in *Agglomeration Economics*, ed. by E. L. Glaeser, National Bureau of Economic Research, 15–65.

CONNELL, J. (1999): "Beyond Manila: Walls, Malls, and Private Spaces," *Environment and Planning A: Economy and Space*, 31, 417–439.

COSTA, D. L. AND M. E. KAHN (2003): "Civic engagement and community heterogeneity: An economist's perspective," *Perspectives on politics*, 1, 103–111, publisher: Cambridge University Press.

COY, M. AND M. PÖHLER (2002): "Gated Communities in Latin American Megacities: Case Studies in Brazil and Argentina," *Environment and Planning B: Planning and Design*, 29, 355–370.

CUNNINGHAM, C. R. (2007): "Growth controls, real options, and land development," *The Review of Economics and Statistics*, 89, 343–358.

DELL, M. AND B. A. OLKEN (2019): "The development effects of the extractive colonial economy: the Dutch cultivation system in Java," *The Review of Economic Studies*, 87, 164–203.

DIPASQUALE, D. AND E. L. GLAESER (1999): "Incentives and social capital: Are homeowners better citizens?" *Journal of urban Economics*, 45, 354–384.

DURANTON, G. (2015): "A proposal to delineate metropolitan areas in Colombia," *Revista Desarrollo y Sociedad*, 223–264.

FREEMAN, L. (2001): "The effects of sprawl on neighborhood social ties: An explanatory analysis," *Journal of the American Planning Association*, 67, 69–77.

GADUH, A. (2016): "Uniter or Divider? Religion and Social Cooperation: Evidence from Indonesia," .

GLAESER, E. L. AND J. D. GOTTLIEB (2006): "Urban Resurgence and the Consumer City," *Urban Studies*, 43, 1275–1299.

GLAESER, E. L., D. I. LAIBSON, AND B. SACERDOTE (2002): "An Economic Approach to Social Capital," *The Economic Journal*, 112, F437–F458.

GLAESER, E. L. AND B. SACERDOTE (1999): "Why Is There More Crime in Cities?" *Journal of Political Economy*, 107, S225–S258.

——— (2000): "The Social Consequences of Housing," *Journal of Housing Economics*, 9, 1 – 23.

GLAESER, E. L. AND B. A. WARD (2009): "The causes and consequences of land use regulation: Evidence from Greater Boston," *Journal of Urban Economics*, 65, 265–278.

GUTMAN, G., C. HUANG, G. CHANDER, P. NOOJIPADY, AND J. G. MASEK (2013): "Assessment of the NASA–USGS Global Land Survey (GLS) datasets," *Remote Sensing of Environment*, 134, 249 – 265.

HABYARIMANA, J., M. HUMPHREYS, D. N. POSNER, AND J. M. WEINSTEIN (2007): "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review*, 101.

HEMANI, S., A. K. DAS, AND A. CHOWDHURY (2017): "Influence of urban forms on social sustainability: A case of Guwahati, Assam," *Urban Design International*, 22, 168–194.

HENGL, T., J. M. DE JESUS, G. B. HEUVELINK, M. R. GONZALEZ, M. KILIBARDA, A. BLAGOTIĆ, W. SHANGGUAN, M. N. WRIGHT, X. GENG, B. BAUER-MARSCHALLINGER, ET AL. (2017): "SoilGrids250m: Global gridded soil information based on machine learning," *PLoS one*, 12.

HILL, H. (1996): *The Indonesian economy since 1966: Southeast Asia's emerging giant*, Cambridge, UK ; New York: Cambridge University Press.

HILL, H., B. P. RESOSUDARMO, AND Y. VIDYATTAMA (2009): "Economic Geography of Indonesia: Location, Connectivity, and Resources," in *Reshaping Economic Geography in East Asia*, ed. by Y. Huang and A. M. Bocchi, Washington, DC: World Bank.

HOXBY, C. M. (2000): "Does competition among public schools benefit students and taxpayers?" *American Economic Review*, 90, 1209–1238.

JACOBS, J. (1961): *The Death and Life of Great American Cities*, New York: Vintage Books.

JELLINEK, L. (1991): *The wheel of fortune: the history of a poor community in Jakarta*, Honolulu: Univ. of Hawaii Press, oCLC: 231222085.

KLEIBERGEN, F. AND R. PAAP (2006): "Generalized reduced rank tests using the singular value decomposition," *Journal of Econometrics*, 133, 97 – 126.

KLING, J. R., J. B. LIEBMAN, AND L. F. KATZ (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75, 83–119.

KOENTJARANINGRAT (1985): *Javanese culture*, Oxford University Press.

LEGATUM INSTITUTE (2019a): *The Legatum Prosperity Index 2019: A tool for transformation*, London: Legatum Institute, 13 ed.

——— (2019b): *The Legatum Prosperity Index 2019: Methodology Report*, London: Legatum Institute, 13 ed.

LEYDEN, K. M. (2003): "Social capital and the built environment: the importance of walkable neighborhoods," *American journal of public health*, 93, 1546–1551.

MAVRIDIS, D. (2015): "Ethnic Diversity and Social Capital in Indonesia," *World Development*, 67, 376–395.

MCGEE, T. G. (1991): "The emergence of desakota regions in Asia: expanding a hypothesis," *The extended metropolis: Settlement transition in Asia*, publisher: University of Hawaii Press.

MUNSHI, K. AND M. ROSENZWEIG (2016): "Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap," *American Economic Review*, 106, 46–98.

MUZAYANAH, I. F. U., S. NAZARA, B. R. MAHI, AND D. HARTONO (2020): "Is there social capital in cities? The association of urban form and social capital formation in the metropolitan cities of Indonesia," *International Journal of Urban Sciences*, 1–25.

NGUYEN, D. (2010): "Evidence of the impacts of urban sprawl on social capital," *Environment and Planning B: Planning and Design*, 37, 610–627.

PESARESI, M., D. EHRLICH, S. FERRI, A. FLORCZYK, S. FREIRE, M. HALKIA, A. JULEA, T. KEMPER, P. SOILLE, V. SYRRIS, ET AL. (2016): "Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014," Tech. rep., Publications Office of the European Union.

PICARD, P. M. AND Y. ZENOU (2018): "Urban spatial structure, employment and social ties," *Journal of Urban Economics*, 104, 77 – 93.

PUTNAM, R. D. (1995): "Tuning in, tuning out: The strange disappearance of social capital in America," *PS: Political science & politics*, 28, 664–684.

——— (2000): *Bowling alone: The collapse and revival of American community*, Simon and schuster.

——— (2007): "E Pluribus Unum: Diversity and Community in the Twenty-first Century The 2006 Johan Skytte Prize Lecture," *Scandinavian Political Studies*, 30, 137–174.

ROBERTS, M., F. G. SANDER, AND S. TIWARI (2019): "Time to Act: Realizing Indonesia's Urban Potential," Tech. rep., World Bank.

ROSENTHAL, S. S. AND W. C. STRANGE (2008): "The Attenuation of Human Capital Spillovers," *Journal of Urban Economics*, 64, 373 – 389.

RUKMANA, D. (2015): "The Change and Transformation of Indonesian Spatial Planning after Suharto's New Order Regime: The Case of the Jakarta Metropolitan Area," *International Planning Studies*, 20, 350–370.

SATO, Y. AND Y. ZENOU (2015): "How urbanization affect employment and social interactions," *European Economic Review*, 75, 131–155.

SIMMEL, G. (1903): *The Metropolis and Mental Life*, New York: Free Press.

TOWNSEND, R. M. (1994): "Risk and insurance in village India," *Econometrica: journal of the Econometric Society*, 539–591.

UCHIDA, H. AND A. NELSON (2010): "Agglomeration Index: Towards a New Measure of Urban Concentration," Background paper for World Development Report 2009: Reshaping Economic Geography. World Bank, Washington, DC.

UN STATISTICAL COMMISSION (2020): "A Recommendation on the Method to Delineate Cities, Urban and Rural Areas for International Statistical Comparisons," Background paper, United Nations Statistical Commission.

UNITED NATIONS (UN) (2019): "World Urbanization Prospects: The 2018 Revision," Technical report, United Nations, Department of Economic and Social Affairs.

USDA (1975): *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 436, US Government Print.

——— (2015): *Illustrated Guide to Soil Taxonomy*, National Resources Conservation Service Lincoln, Nebraska.

VAN KIPPERSLUIS, H. AND C. A. RIETVELD (2018): "Beyond plausibly exogenous," *The Econometrics Journal*, 21, 316–331.

WANG, Y. (2021): "Linguistic Distance, Internal Migration and Welfare: Evidence from Indonesia," *Job Market Paper, Syracuse University*.

WILHELM, M. (2011): "The role of community resilience in adaptation to climate change: The urban poor in Jakarta, Indonesia," in *Resilient Cities*, Springer, 45–53.

WIRTH, L. (1938): "Urbanism as a Way of Life," *American journal of sociology*, 44, 1–24.

WOOD, L., B. GILES-CORTI, AND M. BULSARA (2012): "Streets apart: Does social capital vary with neighbourhood design?" *Urban Studies Research*, 2012.

WOOD, L., T. SHANNON, M. BULSARA, T. PIKORA, G. MCCORMACK, AND B. GILES-CORTI (2008): "The anatomy of the safe and social suburb: an exploratory study of the built environment, social capital and residents' perceptions of safety," *Health & place*, 14, 15–31.

WOOLCOCK, M. AND D. NARAYAN (2000): "Social Capital: Implications for Development Theory, Research, and Policy," *The World Bank Research Observer*, 15, 225–249.

ZHAO, P. (2013): "The Impact of Urban Sprawl on Social Segregation in Beijing and a Limited Role for Spatial Planning," *Tijdschrift voor economische en sociale geografie*, 104, 571–587.

Table 1: Summary Statistics: Social Capital Outcomes

Panel A: Trust in Neighbors	**Description**	**Mean (sd)**	***N***
trust neighbor to watch house	Do you trust your neighbors to watch your house if you are away?	2.91 (0.52)	24,394
trust neighbor to tend children	Do you trust your neighbors to watch your child if there was no adult at home in your house?	2.64 (0.63)	24,394
Panel B: Community Participation	**Description**	**Mean (sd)**	***N***
join community group(s)	Do you usually participate in community activities in the neighborhood (e.g. social gathering, sports, art, etc.)?	2.36 (0.90)	22,790
join religious activities	Do you usually participate in religious activities in the neighborhood (e.g., recitation, religious celebration, etc.)?	2.69 (0.78)	23,967
join religious activities recently	Have you participated in any religious activities in the last 3 months?	0.73 (0.44)	24,105
voluntary public good provision	Do you usually volunteer for your neighborhood (e.g. building public facilities, community service, etc.)?	2.51 (0.82)	23,535
join community activities recently	Have you participated in any community social activities in the last 3 months (e.g. sports, arts, skills dev., funerals, etc.)?	0.80 (0.40)	24,094
Panel C: Social Insurance	**Description**	**Mean (sd)**	***N***
ready to help neighbor	Are you ready to help others who are helpless (need help) in the neighborhood?	2.98 (0.51)	24,394
contribute to assist unfortunate neigbhors	Do you usually help people who are experiencing disasters (such as death, illness, etc.)?	2.81 (0.71)	24,394
easily access to neighbors' help	Is it easy for you to get help from neighbors when you are experiencing financial problems?	2.65 (0.71)	24,394
Panel D: Intergroup Tolerance	**Description**	**Mean (sd)**	***N***
pleased with non-coreligions	How happy are you with activities in the neighborhood by another religion?	2.74 (0.58)	21,659
pleased with non-coethnics	How happy are you with activities in the neighborhood by another ethnic group?	2.82 (0.51)	21,794

Notes: This table reports short titles, longer descriptions, and summary statistics for social capital outcomes from the 2012 *Susenas*. Most of these questions were asked to household head respondents in the social capital module of the *Susenas*, but some were asked in the core module, and for those questions, we only use household head responses. Summary statistics were computed using data from the sample of communities comprising metropolitan areas. The groupings of variables listed here correspond to the groupings used later in the mean effects analysis (e.g., Table 4).

Table 2: First Stage: Log Density vs. Soil Characteristics

	(1)	**(2)**
Soil bulk density at 60 cm depth (kg / m3)	0.039***	0.014***
	(0.003)	(0.002)
Sand content at 60 cm depth (% (kg / kg))	-0.054***	-0.018***
	(0.007)	(0.005)
Great Group: Dystropepts (Inceptisols) (0 1)	-0.708***	-0.227***
	(0.077)	(0.049)
Great Group: Haplustolls (Mollisols) (0 1)	0.933***	0.301***
	(0.150)	(0.086)
Great Group: Haplorthox (Oxisols) (0 1)	-1.015***	-0.520***
	(0.113)	(0.081)
Great Group: Tropudults (Ultisols) (0 1)	-1.351***	-0.809***
	(0.141)	(0.129)
N	2,241	2,241
N Clusters	1,309	1,309
Adj. R^2	0.518	0.775
Adj. R^2 (Within)	0.360	0.701
Regression F-Stat	91.2	99.5
City FE	Yes	Yes
Elevation Control	Yes	Yes
Ruggedness Control	Yes	Yes
Distance to Coast Control	Yes	Yes
Distance to Rivers Control	Yes	Yes
$\mathbf{X}_v$ Controls	.	Yes

Notes: This table reports estimates of equation (6), the community-level first stage relationship between log population density in 2010 (the dependent variable) and different soil characteristics variables. Following Belloni et al. (2014), we use post-double-selection lasso to select the instruments in this regression from a set of 67 soil characteristics. All regressions are limited to the sample of villages within urban areas that appear in the 2012 *Susenas*. We control for city fixed effects, elevation, ruggedness, distance to the nearest point on the coast, and distance to the nearest river. In Column 2, we add the village-level controls for sorting on observables and unobservables, denoted by $\mathbf{X}_v$. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels. A version of this table at the individual level can be found in Appendix Table A.4.

Table 3: The Effect of Density on Trust in Neighbors

	OLS	IV-LASSO
Panel A: Only $\mathbf{X}_i$ and $\mathbf{C}_{2v}$ Controls	**(1)**	**(2)**
Log Density (2010)	-0.026***	-0.056***
	(0.005)	(0.012)
N	23,892	23,892
N Clusters	1,310	1,310
Adjusted R^2	0.034	0.034
Adjusted R^2 (within)	0.011	0.010
Kleibergen-Paap Wald Rank F Stat		82.153
Under Id. Test (KP Rank LM Stat)		221.932
p-Value		0.000
AR Wald Test (Weak IV Robust Inf.)		4.715
p-Value		0.000
Sargan-Hansen Test (Overidentification)		4.245
p-Value		0.515
Panel B: Adding $\mathbf{X}_v$ Controls	**(1)**	**(2)**
Log Density (2010)	-0.007	-0.075***
	(0.009)	(0.026)
N	23,892	23,892
N Clusters	1,310	1,310
Adjusted R^2	0.041	0.041
Adjusted R^2 (within)	0.017	0.018
Kleibergen-Paap Wald Rank F Stat		21.701
Under Id. Test (KP Rank LM Stat)		86.716
p-Value		0.000
AR Wald Test (Weak IV Robust Inf.)		1.866
p-Value		0.083
Sargan-Hansen Test (Overidentification)		2.633
p-Value		0.756
$H_o : \mathbf{\Gamma}_1 = 0$ (p-value)	0.000	0.000
$H_o : \theta_A = \theta_B$ (p-value)	0.061	0.524
City FE	Yes	Yes

Notes: Each cell reports the coefficient on log population density in 2010 from equation (7) where the dependent variable is the 4-point index of trust in neighbors. In the sample, the average of the dependent variable, y, is 2.91, and the standard deviation is 0.514. Column 1 reports OLS estimates, while Column 2 applies a post-double-selection IV-lasso estimator, following Belloni et al. (2012). In Panel A, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. Panel B reports the full, unrestricted model. The specific variables we include in $\mathbf{X}_i$, $\mathbf{C}_{2v}$, and $\mathbf{X}_v$, as well as their coefficients, are reported in Appendix Table A.4 and Appendix Table A.5. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 4: The Effect of Density on Social Capital: Mean Effects

	Trust in Neighbors		Community Participation		Social Insurance		Intergroup Tolerance	
	OLS	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.068***	-0.112***	-0.021***	-0.041***	-0.020***	-0.016	0.055***	0.044**
	(0.008)	(0.017)	(0.005)	(0.010)	(0.006)	(0.014)	(0.010)	(0.021)
2. Adding $\mathbf{X}_v$ Controls	-0.026**	-0.121***	-0.027***	-0.071***	0.014	0.017	0.045***	0.004
	(0.013)	(0.038)	(0.007)	(0.022)	(0.010)	(0.030)	(0.016)	(0.045)
$H_o : \mathbf{\Gamma}_1 = 0$ (p-value)	0.041	0.052	0.000	0.000	0.623	0.595	0.210	0.190
$H_o : \tau_1 = \tau_2$ (p-value)	0.004	0.829	0.474	0.215	0.003	0.315	0.604	0.427
N Outcomes	2	2	5	5	3	3	2	2
N	47,784	47,784	116,144	116,144	71,676	71,676	42,517	42,517
N individuals	23,892	23,892	22,346	22,346	23,892	23,892	21,186	21,186
City FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each cell reports the mean effect estimate, τ, of log population density in 2010 on groups of related outcomes, from equation (8). Outcome groupings are listed in the column headers, and the outcomes themselves are reported in Table 1. Columns 1, 3, 5, and 7 report OLS estimates, while Columns 2, 4, 6, and 8 use post-double-selection IV-Lasso estimates, following Belloni et al. (2012). In row 1, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. Row 2 reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 5: The Effect of Density on Social Capital: Mean Effects (IFLS 5)

	Trust in Neighbors		Community Participation		Social Insurance		Intergroup Tolerance	
	OLS	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.049***	-0.105***	-0.030***	-0.049***	-0.013	-0.051**	0.078***	0.117***
	(0.007)	(0.014)	(0.005)	(0.010)	(0.009)	(0.022)	(0.006)	(0.011)
2. Adding $\mathbf{X}_v$ Controls	-0.006	-0.109**	-0.028***	-0.072***	0.011	-0.127**	0.008	0.045
	(0.008)	(0.045)	(0.006)	(0.028)	(0.013)	(0.064)	(0.006)	(0.027)
$H_o : \mathbf{\Gamma}_1 = 0$ (p-value)	0.907	0.504	0.000	0.000	0.635	0.262	0.000	0.000
$H_o : \tau_1 = \tau_2$ (p-value)	0.000	0.939	0.804	0.443	0.137	0.257	0.000	0.014
N Outcomes	3	3	6	6	1	1	7	7
N	44,102	44,102	97,554	97,554	16,365	16,365	114,553	114,553
N individuals	11,745	11,745	16,259	16,259	16,365	16,365	16,364	16,364
City FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each cell reports the mean effect estimate, τ, of log population density in 2010 on groups of related outcomes, from equation (8), but using the IFLS 5 data for outcomes. Outcome groupings are listed in the column headers, and the outcomes themselves are reported in Appendix Table A.17. Columns 1, 3, 5, and 7 report OLS estimates, while Columns 2, 4, 6, and 8 use post-double-selection IV-Lasso estimates, following Belloni et al. (2012). In row 1, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. Row 2 reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 6: The Effect of Density on Social Capital: IFLS Panel Regressions

	Trust in Neighbors		**Community Participation**		**Social Insurance**		**Intergroup Tolerance**	
	OLS	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Log Density (2010)	-0.074*	-0.063	-0.009	-0.005	-0.094***	-0.227***	0.052**	0.052
	(0.039)	(0.061)	(0.015)	(0.019)	(0.034)	(0.068)	(0.026)	(0.041)
N	169	169	169	169	166	166	169	169
Adjusted R^2	0.038	-0.017	-0.010	-0.035	0.072	-0.079	0.048	-0.014
Kleibergen-Paap Wald Rank F Stat		27.329		27.329		27.748		27.329
Under Id. Test (KP Rank LM Stat)		29.208		29.208		29.196		29.208
p-Value		0.000		0.000		0.000		0.000
AR Wald Test (Weak IV Robust Inf.)		1.149		1.263		4.909		0.923
p-Value		0.331		0.289		0.003		0.431
City FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

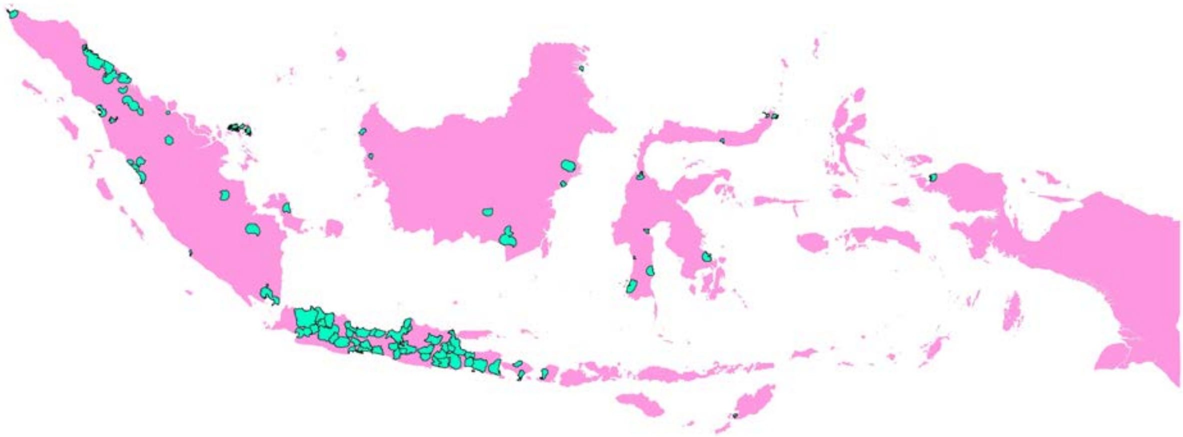
Notes: This table reports mean effect estimates of the impact of density on social capital, using IFLS panel data and a two-step estimation procedure described by Combes et al. (2008) and Combes et al. (2010). In the first step, we use the panel data to estimate local, time-varying effects of social capital after conditioning out the impact of individual-specific effects and the effect of time-varying individual-level observables. We then average the residuals from this regression, and estimate a cross-sectional regression of the average social capital measures (averaged over village years), instrumenting for our density measure in 2010 with the instruments listed in the column headers. See the text for further discussion. Outcome groupings are listed in the column headers. Columns 1, 3, 5, and 7 report OLS estimates, while Columns 2, 4, 6, and 8 use post-double-selection IV-Lasso estimates, following Belloni et al. (2012). All regressions are limited to the sample of IFLS villages within urban areas and include city-fixed effects. Robust standard errors are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table 7: Heterogeneous Effects of Density on Social Capital

	Trust in Neighbors	Community Participation	Social Insurance	Intergroup Tolerance
Panel A: Baseline	**(1)**	**(2)**	**(3)**	**(4)**
Log density	-0.121***	-0.071***	0.017	0.004
	(0.038)	(0.022)	(0.030)	(0.045)
Panel B: Individual-Level Interactions	**(1)**	**(2)**	**(3)**	**(4)**
Log density	-0.109***	-0.059***	0.028	-0.021
	(0.037)	(0.022)	(0.030)	(0.046)
... × Education: High	-0.007	-0.030***	-0.026**	0.024
	(0.015)	(0.010)	(0.013)	(0.018)
Log density	-0.107***	-0.063***	0.008	-0.035
	(0.039)	(0.024)	(0.032)	(0.049)
... × Income: High	-0.011	-0.013	-0.015	0.058***
	(0.016)	(0.010)	(0.012)	(0.020)
Log density	-0.120***	-0.065***	0.015	-0.004
	(0.039)	(0.023)	(0.030)	(0.048)
... × Transportation: Private	-0.026*	-0.020**	-0.027**	0.010
	(0.015)	(0.010)	(0.012)	(0.017)
Panel C: City-Level Interactions	**(1)**	**(2)**	**(3)**	**(4)**
Log density	-0.133***	-0.062***	0.054*	-0.001
	(0.036)	(0.021)	(0.029)	(0.046)
... × Commute Distance: High	-0.013	-0.004	-0.053***	-0.003
	(0.025)	(0.015)	(0.020)	(0.031)
Log density	-0.074*	-0.070***	0.015	0.006
	(0.040)	(0.023)	(0.032)	(0.049)
... × Property Crime: High	-0.062**	-0.029*	-0.027	-0.021
	(0.027)	(0.016)	(0.021)	(0.035)
Log density	-0.142***	-0.048*	0.014	-0.064
	(0.041)	(0.025)	(0.034)	(0.056)
... × Violent Crime: High	0.035	-0.026*	0.015	0.062*
	(0.027)	(0.016)	(0.022)	(0.037)

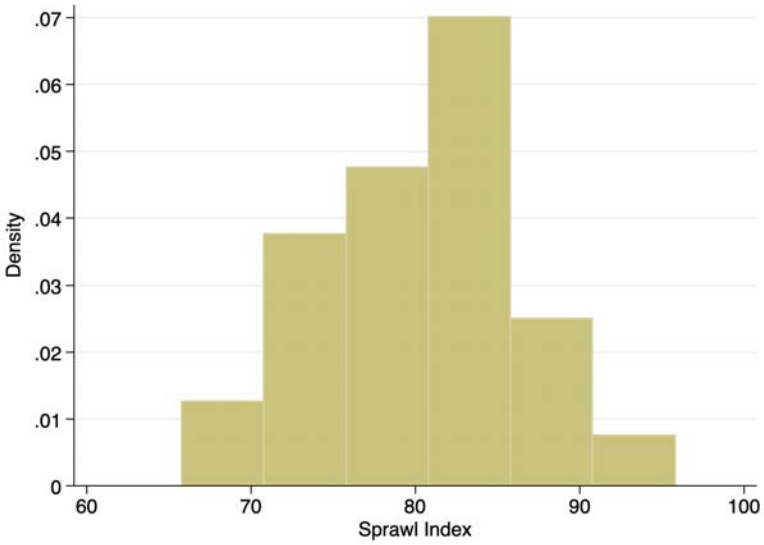
Notes: This table reports mean effects and mean interaction terms for the impact of density on social capital, using the selected soil characteristics from the post-double-selection IV-Lasso procedure and their interactions as instruments and including both the $\mathbf{C}_{2v}$ and $\mathbf{X}_v$ controls. For each panel, Column 1 replicates estimates from Table 4, Column 3, Row 3. See Appendix C.7 for more details on the estimation methodology and variables used for Columns 2-6. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Figure 1: Indonesia's Urban Areas



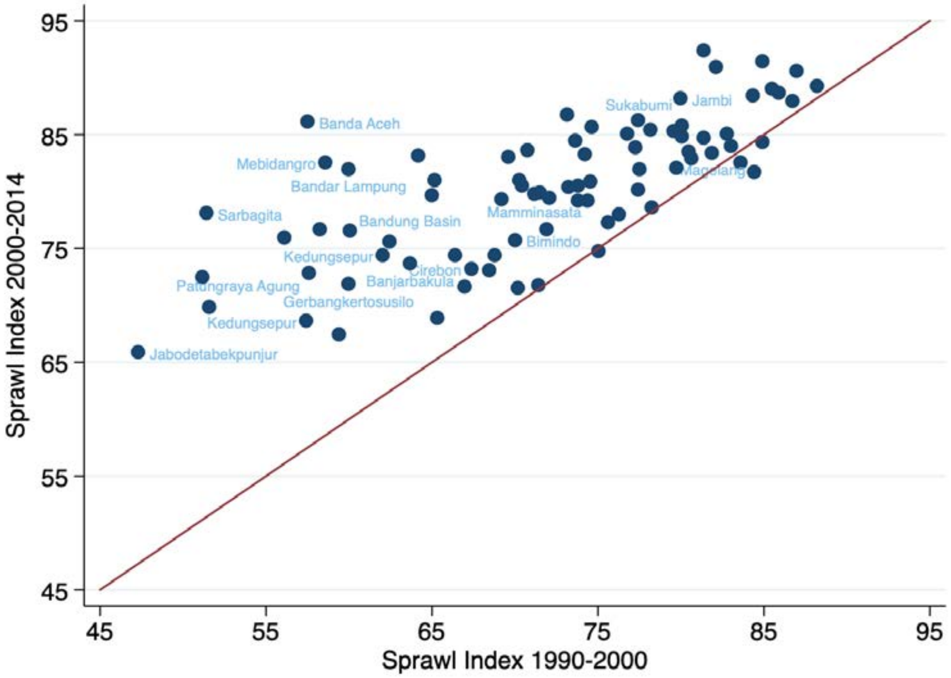
Notes: This figure presents a map of urban areas in Indonesia, where our approach for delineating metro areas, which follows Burchfield et al. (2006), is described in Section 4. We delineated 80 urban metropolitan areas in Indonesia out of the 83 metropolitan areas initially listed by the EAP-UE project. The remaining 3 areas were dropped because they either lacked a well-identified core or did not exhibit sufficiently strong urban expansion in 2014.

Figure 2: Distribution of Urban Sprawl across Indonesian Cities



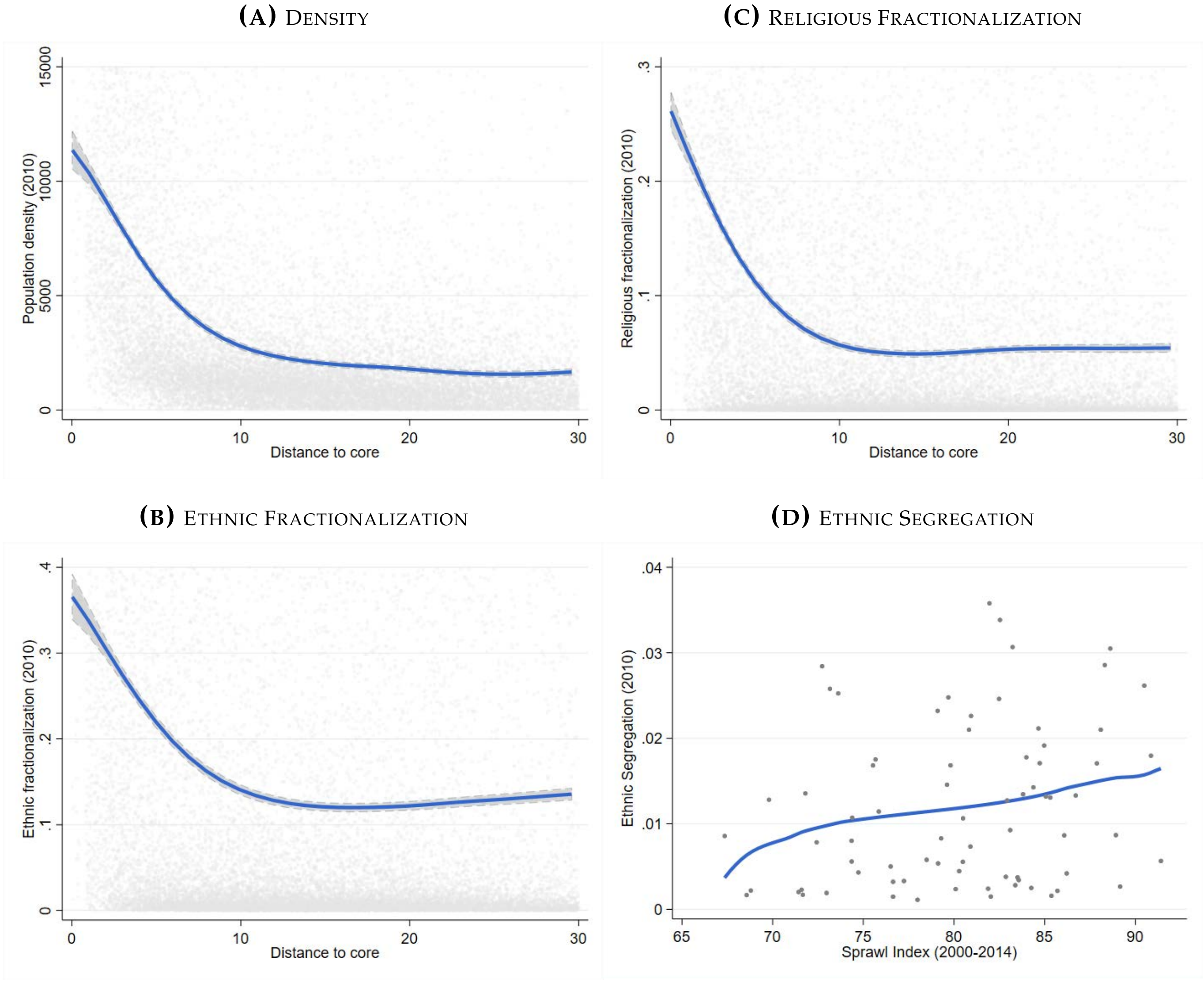
Notes: This is a histogram of the sprawl indices across Indonesian cities, where the sprawl measure is described in Section 4. The sprawl index ranges from a minimum of 65.8 to a maximum of 92.3, with mean of 80.3 and standard deviation equal to 6.1.

Figure 3: Evolution of the Urban Sprawl Indexes over Time



Notes: This figure presents a scatterplot of the relationship between urban sprawl from 2000-2014 against urban sprawl from 1990-2000. Each point represents a different city.

Figure 4: Population Density, Ethnic Composition, and Distance to the CBD



(A) Density

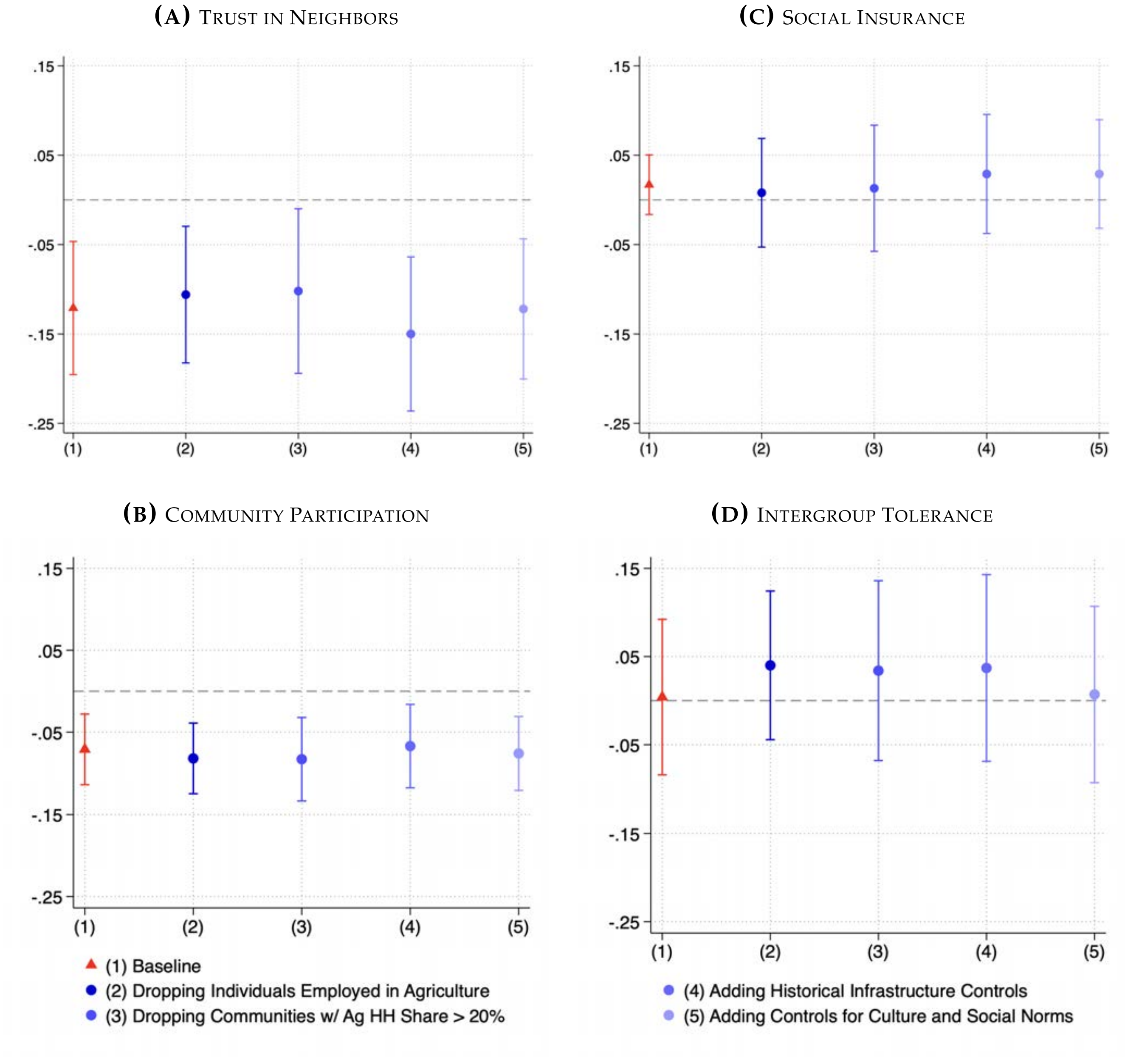
(C) Religious Fractionalization

(B) Ethnic Fractionalization

(D) Ethnic Segregation

Notes: These figures plot local polynomial smooth of census trends on distance to CBD. The smooth uses an Epanechnikov kernel, rule-of-thumb bandwidth and local cubic function. Panel D (Segregation) plots the Alesina and Zhuravskaya (2011) across cities in Indonesia, using villages as the small unit. We omit two cities (Greater Jakarta and Pulang Pisau) from this figure because they are outliers in the relationship.

Figure 5: Robustness of Mean Effects Estimates



(A) Trust in Neighbors

(C) Social Insurance

(B) Community Participation

(D) Intergroup Tolerance

Notes: These figures plot coefficients from mean effect regressions using different specifications described in Section 6.1 in testing the exclusion restrictions of our soil characteristics IVs. Each panel corresponds to a different social capital outcome group, as indicated by the panel title. Within each panel, each dot and band plot the point estimate and 95% confidence interval from a different specification as indicated by the legend. Specification (1) is our baseline specification, from Table 4, row 2. Specification (2) is based on Appendix Table A.10, column 2. Specification (3) is based on Appendix Table A.10, column 6. Specification (4) is based on Appendix Table A.14, column 3. Specification (5) is based on Appendix Table A.14, column 4.

Online Appendix

Civelli, A., Gaduh, A., Rothenberg, A., and Wang, Y. (2021): "Urban Sprawl and Social Capital: Evidence from Indonesian Cities"

Table of Contents

List of Tables

List of Figures

A Additional Tables and Figures

Table A.1: Principal Components Analysis of $\mathbf{X}_v$

	Full ***Susenas***	**Urban** ***Susenas***
	(1)	**(2)**
# of Variables in $\mathbf{X}_v$	38	38
# of factors needed to explain:		
... 75% of total $\mathbf{X}_v$ variation	18	16
... 90% of total $\mathbf{X}_v$ variation	25	23
... 95% of total $\mathbf{X}_v$ variation	28	27
... 99% of total $\mathbf{X}_v$ variation	33	32
... 100% of total $\mathbf{X}_v$ variation	38	37

Notes: This table reports a principal components analysis of the 38 $\mathbf{X}_v$ variables, both for the full *Susenas* sample (column 1) and for our urban *Susenas* sample described in Section 4 (column 2). The first row lists the number of variables in $\mathbf{X}_v$. The next set of rows report the number of factors needed to explain 75%, 90%, 95%, 99% and 100% of the total variation in $\mathbf{X}_v$.

Table A.2: Kleibergen and Paap (2006) Cluster-Robust Tests of the Rank of the $\mathbf{X}_v$ Covariance Matrix

# Fact.		Full *Susenas*	Urban *Susenas*
		P-Value	P-Value
H_0	H_A	(1)	(2)
10	11+	0.000	0.008
11	12+	0.000	0.003
12	13+	0.000	0.000
13	14+	0.000	0.000
14	15+	0.000	0.000
15	16+	0.000	0.000
16	17+	0.000	0.000
17	18+	0.000	0.000
18	19+	0.000	0.000
19	20+	0.000	0.000
20	21+	0.000	0.000
21	22+	0.000	0.000
22	23+	0.000	0.000
23	24+	0.000	0.000
24	25+	0.000	0.000
25	26+	0.000	0.000
26	27+	0.000	0.002
27	28+	0.000	0.072
28	29+	0.000	0.217
29	30+	0.000	0.683
30	31+	0.000	0.927
31	32+	0.056	0.942
32	33+	0.100	0.962
33	34+	0.253	0.988
34	35+	0.206	0.998
35	36+	0.644	0.996
36	37+	0.717	0.967
37	38+	0.998	0.861

Notes: Each element of this table reports a p-value from a test based on Kleibergen and Paap (2006) of the null hypothesis that the rank of the covariance matrix of $\mathbf{X}_v$ is equal to the value associated with the row label, against the alternative that the rank exceeds this value. These p-values are robust and account for clustering at the sub-district level. Column 1 performs these tests on the full *Susenas* sample, while column 2 performs them just for our urban *Susenas* sample, described in Section 4.

Table A.3: Individual-Level First Stage: Log Density vs. Soil Characteristics

	(1)	(2)
Soil bulk density at 60 cm depth (kg / m3)	0.034***	0.014***
	(0.002)	(0.002)
Sand content at 60 cm depth (% (kg / kg))	-0.045***	-0.018***
	(0.007)	(0.005)
Great Group: Dystropepts (Inceptisols) (0 1)	-0.595***	-0.218***
	(0.069)	(0.050)
Great Group: Haplustolls (Mollisols) (0 1)	0.744***	0.305***
	(0.139)	(0.088)
Great Group: Haplorthox (Oxisols) (0 1)	-0.949***	-0.558***
	(0.107)	(0.084)
Great Group: Tropudults (Ultisols) (0 1)	-1.259***	-0.831***
	(0.132)	(0.124)
N	23,942	23,942
N Clusters	1,310	1,310
Adj. R^2	0.593	0.792
Adj. R^2 (Within)	0.434	0.710
Regression F-Stat	76.3	82.9
City FE	Yes	Yes
Elevation Control	Yes	Yes
Ruggedness Control	Yes	Yes
Distance to Coast Control	Yes	Yes
Distance to Rivers Control	Yes	Yes
$\mathbf{X}_i$ Controls	Yes	Yes
$\mathbf{X}_v$ Controls	.	Yes

Notes: This table reports estimates of the first stage relationship between log population density in 2010 (the dependent variable) and different soil characteristics variables from individual-level *Susenas* specifications. The sample includes all individuals from the *Susenas* within urban areas who have responded to the trust in neighbors question (the dependent variable in Table 3). Following Belloni et al. (2014), we use post-double-selection lasso to select the instruments in this regression from a set of 67 soil characteristics. All regressions include city-fixed effects, individual-level controls, $\mathbf{X}_i$, and controls for elevation, ruggedness, distance to the nearest point on the coast, and distance to the nearest river. In Column 2, we add the village-level controls for sorting on observables and unobservables, denoted by $\mathbf{X}_v$. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.4: The Effect of Density on Trust in Neighbors: Individual Controls

	OLS	IV-LASSO
Panel A: Only $\mathbf{X}_i$ and $\mathbf{C}_{2v}$ Controls	**(1)**	**(2)**
Log Density (2010)	-0.026***	-0.056***
	(0.005)	(0.012)
Age	0.004**	0.004**
	(0.002)	(0.002)
Age^2	-0.000*	-0.000**
	(0.000)	(0.000)
Female (0 1)	-0.009	-0.009
	(0.016)	(0.016)
Some High School (0 1)	-0.018*	-0.010
	(0.010)	(0.010)
Only Completed High School (0 1)	-0.058***	-0.041***
	(0.013)	(0.015)
Any Higher Education (0 1)	-0.075***	-0.059***
	(0.017)	(0.018)
Married (0 1)	0.006	-0.008
	(0.024)	(0.024)
Divorced (0 1)	0.005	-0.008
	(0.030)	(0.031)
Widowed (0 1)	0.010	-0.001
	(0.026)	(0.026)
Working in Agriculture (0 1)	0.028**	0.006
	(0.011)	(0.014)
Non-employed (0 1)	-0.016	-0.016
	(0.014)	(0.014)
Self-employed (0 1)	-0.005	-0.000
	(0.010)	(0.010)
Employer (0 1)	-0.030***	-0.029***
	(0.011)	(0.011)
Household Size	-0.006**	-0.005**
	(0.002)	(0.002)
Ever Migrant (0 1)	-0.011	-0.003
	(0.011)	(0.011)
Recent Migrant (0 1)	-0.006	-0.008
	(0.023)	(0.022)
Elevation	-0.000	-0.000**
	(0.000)	(0.000)
Ruggedness	-0.066*	-0.091**
	(0.038)	(0.039)
Distance to the coast	0.009	0.004
	(0.006)	(0.007)
Distance to nearest river	0.001	-0.003
	(0.005)	(0.005)
N	23,892	23,892
N Clusters	1,310	1,310
Adjusted R^2	0.034	0.034
Adjusted R^2 (within)	0.011	0.010
Kleibergen-Paap Wald Rank F Stat		82.153
Under Id. Test (KP Rank LM Stat)		221.932
p-Value		0.000
AR Wald Test (Weak IV Robust Inf.)		4.715
p-Value		0.000
Sargan-Hansen Test (Overidentification)		4.245
p-Value		0.515
$\mathbf{C}_{2v}$ Controls	Yes	Yes
City FE	Yes	Yes

Notes: This table reports estimates of θ and β from equation (7), where we set $\mathbf{\Gamma}_1 = 0$. This specification is identical to Table 3, Panel A, but we report estimates of β here instead of supressing them as we do in the main table. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.5: The Effect of Density on Trust in Neighbors: Adding $\mathbf{X}_v$

	OLS	**IV-LASSO**
Panel B: Adding $\mathbf{X}_v$ Controls	**(1)**	**(2)**
Log Density (2010)	-0.007	-0.075***
	(0.009)	(0.026)
Avg. Age	-0.004	-0.008
	(0.007)	(0.007)
Percent Female	0.087	-0.051
	(0.495)	(0.503)
Percent Working in Agriculture	0.054	-0.129
	(0.056)	(0.085)
Avg. Household Size	-0.000	-0.000
	(0.000)	(0.000)
Avg. Years of Schooling	-0.027**	-0.006
	(0.011)	(0.014)
Percent Single	-1.094	-1.527
	(1.085)	(1.075)
Percent Married	-0.738	-1.231
	(0.997)	(0.996)
Percent Divorced	-2.573	-2.515
	(2.006)	(2.028)
Percent Unemployed	-0.077	-0.105
	(0.126)	(0.126)
Percent Self-employed	-0.171	-0.110
	(0.192)	(0.190)
Percent Employer	-0.315	-0.212
	(0.207)	(0.212)
Percent Ever Migrants	0.097	0.149**
	(0.072)	(0.075)
Percent Recent Migrants	-0.112	-0.326*
	(0.155)	(0.168)
Percent Speak Indonesian	0.277*	0.142
	(0.165)	(0.167)
Percent Religion: Islam	0.474	1.087
	(1.662)	(1.655)
Percent Religion: Christian	0.543	1.153
	(1.667)	(1.660)
Percent Religion: Catholic	0.208	0.588
	(1.691)	(1.668)
Percent Religion: Hindu	0.337	0.888
	(1.678)	(1.665)
Percent Religion: Buddhist	0.621	1.168
	(1.670)	(1.656)
Percent Religion: Confucian	0.242	0.137
	(1.717)	(1.686)
Percent Jawa	-0.010	0.005
	(0.076)	(0.075)
Percent Sunda	0.020	0.013
	(0.072)	(0.071)
Percent Batak	-0.013	0.063
	(0.135)	(0.139)
Percent Ethnicities from Nusa Tenggara	0.064	0.040
	(0.303)	(0.289)
Percent Madura	-0.114	-0.037
	(0.148)	(0.147)
Percent Betawi	-0.069	-0.054
	(0.148)	(0.147)
Percent Aceh	-1.153*	-1.324*
	(0.690)	(0.725)
Percent Minangkabau	0.351	0.674*
	(0.305)	(0.351)
Percent Bugis	-0.325	-0.317
	(0.273)	(0.303)
Percent Malay	-0.225	-0.124
	(0.187)	(0.185)
Percent Ethnicities from South Sumatra	-0.339*	-0.265
	(0.177)	(0.192)
Percent Ethnicities from Banten	-0.614**	-0.646**
	(0.297)	(0.288)
Percent Banjar	-0.005	0.115
	(0.225)	(0.240)
Percent Dayak	-0.291	-0.255
	(0.392)	(0.403)
Percent Chinese	-0.445*	-0.196
	(0.263)	(0.273)
Percent Ethnicities from Central Sulawesi	0.051	-0.055
	(0.240)	(0.260)
Percent Ethnicities from Papua	-6.853	-6.280
	(4.583)	(4.588)
Percent Makassar	-0.148	-0.161
	(0.270)	(0.291)
N	23,892	23,892
N Clusters	1,310	1,310
Adjusted R^2	0.041	0.041
Adjusted R^2 (within)	0.017	0.018
Kleibergen-Paap Wald Rank F Stat		21.701
Under Id. Test (KP Rank LM Stat)		86.716
p-Value		0.000
AR Wald Test (Weak IV Robust Inf.)		1.866
p-Value		0.083
Sargan-Hansen Test (Overidentification)		2.633
p-Value		0.756
$\mathbf{X}_i$ Controls	Yes	Yes
$\mathbf{C}_{2v}$ Controls	Yes	Yes
City FE	Yes	Yes

Notes: This table reports estimates of θ and $\mathbf{\Gamma}$ from equation (7). This specification is identical to Table 3, Panel B, but we report estimates of $\mathbf{\Gamma}$ here instead of supressing them as we do in the main table. Estimates of β are supressed. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.6: The Effect of Density on Trust in Neighbors and Community Participation (Linear Index)

	OLS (1)	IV-LASSO (2)	Dep.Var Mean (SE)	N
Panel A: Trust in Neighbors				
1. trust neighbor to watch house	-0.026*** (0.005)	-0.056*** (0.012)	2.916 (0.003)	23,892
... adding $\mathbf{X}_v$ controls	-0.006 (0.009)	-0.075*** (0.026)		
2. trust neighbor to tend children	-0.054*** (0.007)	-0.072*** (0.016)	2.648 (0.004)	23,892
... adding $\mathbf{X}_v$ controls	-0.022** (0.011)	-0.061* (0.035)		
Panel B: Community Participation				
3. join community group(s)	0.015 (0.009)	-0.003 (0.018)	2.365 (0.006)	22,346
... adding $\mathbf{X}_v$ controls	-0.011 (0.013)	-0.063 (0.043)		
4. join religious activities	-0.031*** (0.008)	-0.037** (0.016)	2.689 (0.005)	23,498
... adding $\mathbf{X}_v$ controls	-0.024** (0.011)	-0.033 (0.036)		
5. join religious activities recently	-0.015*** (0.004)	-0.034*** (0.010)	0.730 (0.003)	23,616
... adding $\mathbf{X}_v$ controls	-0.015** (0.007)	-0.045** (0.022)		
6. voluntary public good provision	-0.022*** (0.008)	-0.033* (0.019)	2.507 (0.005)	23,081
... adding $\mathbf{X}_v$ controls	-0.012 (0.013)	-0.042 (0.043)		
7. join community activities recently	-0.008* (0.005)	-0.016* (0.009)	0.798 (0.003)	23,603
... adding $\mathbf{X}_v$ controls	-0.010 (0.007)	-0.035* (0.020)		
City FE	Yes	Yes		

Notes: Each cell reports the coefficient on log population density in 2010 from a separate linear regression of (7) where the dependent variable is the outcome listed in the row header. Column 1 reports OLS estimates, while Column 2 applies a post-double-selection lasso estimator to select the best soil characteristics instruments, following Belloni et al. (2012). In the first row of each panel, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. The second row of each panel reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.7: The Effect of Density on Social Insurance and Intergroup Tolerance (Linear Index)

	OLS (1)	IV-LASSO (2)	Dep.Var Mean (SE)	N
Panel C: Social Insurance				
8. ready to help neighbor	0.002 (0.006)	0.016 (0.011)	2.981 (0.003)	23,892
... adding $\mathbf{X}_v$ controls	0.017** (0.008)	0.040 (0.025)		
9. contribute to assist unfortunate neigbhors	-0.019** (0.007)	-0.016 (0.016)	2.809 (0.005)	23,892
... adding $\mathbf{X}_v$ controls	-0.002 (0.011)	0.005 (0.035)		
10. easily access to neighbors' help	-0.027*** (0.008)	-0.041** (0.018)	2.653 (0.005)	23,892
... adding $\mathbf{X}_v$ controls	0.019 (0.012)	-0.025 (0.039)		
Panel D: Intergroup Tolerance				
11. pleased with non-coreligions	0.046*** (0.009)	0.042** (0.018)	2.736 (0.004)	21,186
... adding $\mathbf{X}_v$ controls	0.037*** (0.013)	0.027 (0.038)		
12. pleased with non-coethnics	0.016** (0.007)	0.008 (0.014)	2.822 (0.004)	21,331
... adding $\mathbf{X}_v$ controls	0.014 (0.012)	-0.019 (0.032)		
City FE	Yes	Yes		

Notes: Each cell reports the coefficient on log population density in 2010 from a separate linear regression of (7) where the dependent variable is the outcome listed in the row header. Column 1 reports OLS estimates, while Column 2 applies a post-double-selection lasso estimator to select the best soil characteristics instruments, following Belloni et al. (2012). In the first row of each panel, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. The second row of each panel reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Relationship Between Agricultural Productivity and Selected Soil Characteristics. To assess whether our selected soil characteristics IVs predicted agricultural productivity in urban areas, we began by constructing revenue-weighted log productivity measures for different crops. To measure agricultural productivity, we used yields data from the 2002 Indonesian Village Potential Survey (or *Podes*) and national crop prices from FAO/PriceStat data.[45] The crop categories includes: (1) rice; (2) secondary food crops, known collectively as *palawija*, which include maize, cassava, groundnuts, sweet potato, and soybeans; (3) cash crops, the most important of which are palm oil, rubber, cocoa, and coffee; and (4) total agricultural production, which includes all crops. The crop productivity measures are recorded in log revenue-weighted yield (tons) per hectare.

Appendix Table A.8 shows that conditional on urban area fixed effects, community characteristics, and the sorting controls, the selected soil characteristics that we use to predict population density are not individually significant in predicting rice productivity, food crop productivity, cash crop productivity, or total agricultural productivity. Although soil bulk density is weakly predictive of variation in secondary food crop productivity, as a whole the variables are not jointly significant.

Table A.8: The Effect of Soil Characteristics on Agricultural Productivity in Urban Areas

	Log Rice Productivity	**Log Food Crop Productivity**	**Log Cash Productivity**	**Log Total Productivity**
	(1)	**(2)**	**(3)**	**(4)**
Soil bulk density at 60 cm depth (kg / m3)	-0.002	-0.008*	0.002	-0.001
	(0.003)	(0.005)	(0.006)	(0.003)
Sand Content in H2O at 60 cm depth (% (kg / kg))	0.002	0.008	-0.004	0.007
	(0.008)	(0.013)	(0.015)	(0.007)
Great Group: Dystropepts (Inceptisols) (0 1)	0.044	0.140	-0.020	-0.007
	(0.061)	(0.085)	(0.130)	(0.066)
Great Group: Haplustolls (Mollisols) (0 1)	0.049	-0.049	0.063	0.124
	(0.138)	(0.181)	(0.422)	(0.169)
Great Group: Haplorthox (Oxisols) (0 1)	0.121	0.119	-0.148	-0.061
	(0.080)	(0.141)	(0.222)	(0.098)
Great Group: Tropudults (Ultisols) (0 1)	-0.139	0.138	-0.205	-0.110
	(0.118)	(0.214)	(0.225)	(0.144)
N	1,625	1,074	1,036	1,625
N Clusters	1,045	747	729	1,045
Adj. R^2	0.194	0.168	0.154	0.185
Adj. R^2 (Within)	0.047	0.002	0.020	0.013
Regression F-Stat	2.8	1.6	1.8	1.2
$H_o : \beta_1 = 0$ (p-value)	0.542	0.377	0.979	0.831
Control $\mathbf{C}_{2v}$	Yes	Yes	Yes	Yes
Control $\mathbf{X}_v$	Yes	Yes	Yes	Yes
City FE	Yes	Yes	Yes	Yes

Notes: This table reports linear regression estimates of the relationship between different measures of agricultural productivity and the IV-Lasso selected soil characteristics (from Table 2). The dependent variables, listed in the column headers, are measured in log revenue-weighted yield (tons) per hectare. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

[45] The *Podes* is a census of Indonesian villages conducted approximately every three years by BPS. It collects detailed information from community informants about community characteristics, such as demographics, geography, as well as social and economic infrastructure.

Lasso Predictions of Agricultural Productivity with Soil Characteristics. Next, we used the lasso procedure on the full set of soil characteristics to try to predict agricultural productivity in urban areas. Appendix Table A.9 shows that we are unable to select any soil characteristics to predict rice production, cash production, or total agricultural production. For food crop production, a single soil type, was predictive, but this variable was not selected to predict population density. Moreover, the implied Kleibergen-Paap Wald F-stat is only 9.2. In general, we are not very successful in predicting agricultural production with soil characteristics, and when we are, the variables that are selected are not the ones we use as instruments for density.

Table A.9: Selected IVs: Density and Agricultural Productivity in Urban Areas

Specification No.	Independent Var.	# of IVs	Names of IVs	Kleibergen-Paap Wald Rank F Stat
(1)	Log Population Density	6	Soil bulk density at 60 cm depth, Sand content at 60 cm depth, Great Group 255: Dystropepts (Inceptisols) (0 1), Great Group 291: Haplustolls (Mollisols) (0 1), Great Group 340: Haplorthox (Oxisols) (0 1), Great Group 402: Tropudults (Ultisols) (0 1)	21.7
(2)	Log Rice Productivity (Weighted)	0	.	.
(3)	Log Food Crop Productivity (Weighted)	1	Great Group 262:Tropaquepts (Inceptisols)	9.2
(4)	Log Cash Productivity (Weighted)	0	.	.
(5)	Log Total Productivity (Weighted)	0	.	.

Notes: This table reports the soil characteristics variables that were selected to predict the following variables: (1) log population density; (2) log rice productivity; (3) log food crop productivity; (4) log cash crop productivity; and (5) log total agricultural productivity. The four crop productivity measures are recorded in log revenue-weighted yield (tons) per hectare. For this table, we used post-double-selection lasso techniques to select soil characteristics that are the best predictors from a set of 67 variables, following Belloni et al. (2012). The six selected soil characteristics variables for log density correspond to those used in our first stage relationship, reported in Table 2. All specifications include urban-area fixed effects and control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$.

Dropping Agricultural Households. In Table A.10, we show that our mean effects results are robust to excluding agricultural households and communities. In this table, separate panels are used to denote different outcome groupings. Column 1 reproduces our baseline IV-Lasso estimates using all households in the sample, from Table 4. In column 2, we drop all individuals from the sample who are employed in agriculture, using data on employment from the Susenas.

In the next four columns, we used 2010 census data to define agricultural households as those where all employed members report working in agriculture. We exclude communities from the sample where the agricultural household share is over 80 percent (column 3), over 60 percent (column 4), over 40 percent (column 5), and over 20 percent (column 6). The magnitude and significance of the estimated effects of density remain largely robust to excluding these partially agricultural communities. This suggests that agricultural areas in the periphery are not driving our main results.

Table A.10: Mean Effects of Density on Social Capital: Dropping Agricultural Households

	Baseline IV-Lasso	Dropping Individuals Employed in Agriculture	Dropping Communities with Agricultural Household Share			
			>80%	>60%	>40%	>20%
Panel A: Trust in Neighbors	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.112***	-0.105***	-0.113***	-0.118***	-0.113***	-0.107***
	(0.017)	(0.017)	(0.017)	(0.018)	(0.019)	(0.022)
2. Adding $\mathbf{X}_v$ Controls	-0.121***	-0.106***	-0.124***	-0.126***	-0.118***	-0.102**
	(0.038)	(0.039)	(0.038)	(0.038)	(0.040)	(0.047)
N Outcomes	2	2	2	2	2	2
N	47,784	38,390	47,592	46,764	44,504	39,508
N individuals	23,892	19,195	23,796	23,382	22,252	19,754
Panel B: Community Participation	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.041***	-0.046***	-0.044***	-0.044***	-0.044***	-0.044***
	(0.010)	(0.010)	(0.012)	(0.012)	(0.012)	(0.012)
2. Adding $\mathbf{X}_v$ Controls	-0.071***	-0.082***	-0.071***	-0.068***	-0.080***	-0.083***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.023)	(0.026)
N Outcomes	5	5	5	5	5	5
N	116,144	93,269	115,673	113,697	108,314	96,167
N individuals	22,346	17,988	22,256	21,896	20,909	18,599
Panel C: Social Insurance	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.016	-0.024*	-0.016	-0.016	-0.016	-0.016
	(0.013)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)
2. Adding $\mathbf{X}_v$ Controls	0.017	0.008	0.017	0.015	0.015	0.013
	(0.030)	(0.031)	(0.030)	(0.030)	(0.032)	(0.036)
N Outcomes	3	3	3	3	3	3
N	71,676	57,585	71,388	70,146	66,756	59,262
N individuals	23,892	19,195	23,796	23,382	22,252	19,754
Panel D: Intergroup Tolerance	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	0.044**	0.058***	0.045**	0.040*	0.042*	0.060**
	(0.021)	(0.021)	(0.021)	(0.022)	(0.025)	(0.027)
2. Adding $\mathbf{X}_v$ Controls	0.004	0.040	0.004	-0.004	0.018	0.034
	(0.045)	(0.043)	(0.046)	(0.046)	(0.049)	(0.052)
N Outcomes	2	2	2	2	2	2
N	42,517	34,665	42,365	41,643	39,805	35,920
N individuals	21,186	17,280	21,110	20,753	19,833	17,905
City FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each cell reports the mean effect estimate, τ, of log population density in 2010 on groups of related outcomes, from equation (8). Outcome groupings are listed in the panel headers, and the outcomes themselves are reported in Table 1. All columns use post-double-selection IV-Lasso estimates, following Belloni et al. (2012). The first column reproduces our baseline IV-Lasso estimates from Table 4. For each panel, in row 1, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$, while row 2 reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Varying Distance to the Core. In Appendix Table A.11, we also dropped peripheral communities in the sample, as a further way of exploring the sensitivity of our results to including communities far from the city center. We first calculated the distance between each community and the centroid of the city's CBD. We drop communities that are more than 30 kilometers from the CBD (column 2), more than 25 kilometers from the CBD (column 3), and more than 20 kilometers from the CBD (column 4). Overall, the estimated effects of density in both panels are largely robust to dropping these peripheral communties from the sample.

Table A.11: Mean Effects of Density on Social Capital: Varying Distance to the Core

	Baseline	**≤30km**	**≤25km**	**≤20km**
Panel A: Trust in Neighbors	**(1)**	**(2)**	**(3)**	**(4)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.112***	-0.116***	-0.125***	-0.126***
	(0.017)	(0.019)	(0.021)	(0.022)
2. Adding $\mathbf{X}_v$ Controls	-0.121***	-0.129***	-0.143***	-0.127***
	(0.038)	(0.044)	(0.046)	(0.046)
N Outcomes	2	2	2	2
N	47,784	41,798	38,894	34,858
N individuals	23,892	20,899	19,447	17,429
Panel B: Community Participation	**(1)**	**(2)**	**(3)**	**(4)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.041***	-0.039***	-0.040***	-0.034***
	(0.010)	(0.011)	(0.012)	(0.013)
2. Adding $\mathbf{X}_v$ Controls	-0.071***	-0.060**	-0.062**	-0.065**
	(0.022)	(0.025)	(0.026)	(0.027)
N Outcomes	5	5	5	5
N	116,144	101,667	94,622	84,759
N individuals	22,346	19,605	18,266	16,377
Panel C: Social Insurance	**(1)**	**(2)**	**(3)**	**(4)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.016	-0.010	-0.015	-0.026
	(0.013)	(0.016)	(0.017)	(0.018)
2. Adding $\mathbf{X}_v$ Controls	0.017	0.038	0.024	-0.012
	(0.030)	(0.035)	(0.036)	(0.037)
N Outcomes	3	3	3	3
N	71,676	62,697	58,341	52,287
N individuals	23,892	20,899	19,447	17,429
Panel D: Intergroup Tolerance	**(1)**	**(2)**	**(3)**	**(4)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	0.044**	0.036	0.033	0.036
	(0.021)	(0.023)	(0.024)	(0.026)
2. Adding $\mathbf{X}_v$ Controls	0.004	-0.006	-0.003	0.012
	(0.045)	(0.047)	(0.048)	(0.050)
N Outcomes	2	2	2	2
N	42,517	37,444	35,026	31,568
N individuals	21,186	18,663	17,446	15,731
City FE	Yes	Yes	Yes	Yes

Notes: Each cell reports the mean effect estimate, τ, of log population density in 2010 on groups of related outcomes, from equation (8). Outcome groupings are listed in the panel headers, and the outcomes themselves are reported in Table 1. All columns use post-double-selection IV-Lasso estimates, following Belloni et al. (2012). The first column reproduces our baseline IV-Lasso estimates from Table 4. For each panel, in row 1, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\boldsymbol{\Gamma}_1 = 0$, while row 2 reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Column 2 drops communities that are more than 30 kilometers from their city's CBD. Column 3 drops communities that are more than 25 kilometers from their city's CBD. Column 4 drops communities that are more than 20 kilometers from their city's CBD. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Historical Infrastructure. To control for historical infrastructure that could be correlated with soil characteristics and independently affect social capital, we constructed variables from the 1983 *Podes* that measure several aspects of infrastructure investments and community organizations. These variables include: (1) education facilities (the number of kindergarten, primary, junior secondary, and senior-secondary schools); (2) medical facilities (the number of hospitals, the number of community health clinics or *Puskesmas*, and the number of community based preventative and promotive care facilities or *Posyandu*); (3) the number of places of worships (counts of the number of mosques, surau, churches, pura, and vihara); (4) irrigation infrastructure (the share of wet/paddy rice fields that use man-made irrigation); (5) utilities (share of households covered by the national electricity grid); (6) the share of communities with various agricultural and social organizations.[46]

Although such measures were recorded more than 25 years before our social capital outcomes were recorded, an issue is that because village names are not available in the original survey, we cannot merge the 1986 *Podes* data at the village level. Instead, we have to aggregate to the sub-district level before merging, and we lose a few observations because of difficulties in merging sub-districts over a nearly 30 year period. Finally, we also constructed a control for distance to major roads, based on data from 1990 from Indonesia's Ministry of Public Works and Housing.

Appendix Table A.12 shows that including each of these controls separately (and together in the final column) does not change the results of our baseline trust in neighbors regressions (from Table 3). Appendix Table A.14 shows that our mean effects results are also robust to including these controls. Columns 1 and 2 reproduce the baseline OLS and IV-Lasso estimates of the effects of density on social capital, updated for the set of 2012 Susenas household observations that we can successfully merge to the 1983 *Podes*. Column 3 presents the IV-Lasso estimates after adding all of the controls for historical and social infrastructure listed above. Overall, the estimated effects of density, both with and without controls for contemporary sorting, are largely robust to including these additional controls.

[46] Agricultural organization controls include the number of water users groups, intensification groups, rural listeners groups, agricultural women's groups, young farmers groups, and the number of farmers contact groups. Social organization controls include the number of scouting groups, sporting groups, martial arts groups, theater groups, dance groups, and youth associations.

Table A.12: The Effect of Density on Trust in Neighbors: Controlling for Historical Infrastructure

	Baseline		**Control Hist. Infras.**							
	OLS	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**
Panel A: Only $\mathbf{X}_i$ and $\mathbf{C}_{2v}$ Controls	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
Log Density (2010)	-0.031***	-0.064***	-0.062***	-0.062***	-0.063***	-0.074***	-0.064***	-0.075***	-0.073***	-0.089***
	(0.005)	(0.014)	(0.016)	(0.015)	(0.015)	(0.021)	(0.016)	(0.019)	(0.018)	(0.025)
N	22,010	22,010	22,010	22,010	22,010	22,010	22,010	22,010	22,010	22,010
N Clusters	1,221	1,221	1,221	1,221	1,221	1,221	1,221	1,221	1,221	1,221
Adjusted R^2	0.036	0.035	0.042	0.037	0.037	0.037	0.036	0.038	0.035	0.044
Kleibergen-Paap Wald Rank F Stat		79.112	67.009	74.639	74.001	43.230	64.892	52.756	65.053	37.178
Panel B: Adding $\mathbf{X}_v$ Controls	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
Log Density (2010)	-0.013	-0.084***	-0.082***	-0.082***	-0.083***	-0.090***	-0.082***	-0.094***	-0.087***	-0.098***
	(0.009)	(0.030)	(0.029)	(0.029)	(0.030)	(0.032)	(0.030)	(0.031)	(0.031)	(0.033)
N	22,010	22,010	22,010	22,010	22,010	22,010	22,010	22,010	22,010	22,010
N Clusters	1,221	1,221	1,221	1,221	1,221	1,221	1,221	1,221	1,221	1,221
Adjusted R^2	0.041	0.042	0.047	0.043	0.043	0.042	0.042	0.043	0.042	0.049
Kleibergen-Paap Wald Rank F Stat		22.864	23.359	24.122	22.563	20.778	23.691	22.571	24.594	23.298
Historical Infrastructure Controls										
Education Facilities (1983)	.	.	Yes	.	.	.	.	.	.	Yes
Historical Medical Facilities (1983)	.	.	.	Yes	.	.	.	.	.	Yes
Places of Worship(1983)	.	.	.	.	Yes	.	.	.	.	Yes
Irrigation and PLN (1983)	.	.	.	.	.	Yes	.	.	.	Yes
Agricultural Organizations (1983)	.	.	.	.	.	.	Yes	.	.	Yes
Social Activities (1983)	.	.	.	.	.	.	.	Yes	.	Yes
Distance to Major Road	.	.	.	.	.	.	.	.	Yes	Yes
City FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each cell reports the coefficient on log population density in 2010 from equation (7) where the dependent variable is the 4-point index of trust in neighbors. Column 1 and 2 reproduce the OLS estimates and IV-lasso estimates from Table 3. Notice that the estimates are slightly different due to the imperfect merging to historical infrastructures data. In Panel A, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. Panel B reports the full, unrestricted model. The specific variables we include in $\mathbf{X}_i$, $\mathbf{C}_{2v}$, and $\mathbf{X}_v$, as well as their coefficients, are reported in Appendix Table A.4 and Appendix Table A.5. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. From Column 3-9, we seperately includes different groups of historical infrastructure variables, as idicated from the bottom panel. In Column 10, we include all historical infrastructures. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Culture and Social Norms. Yet another concern is that differences in culture that are correlated with soil characteristics could be playing a role in explaining our results. For example, Alesina et al. (2013) show that soil characteristics historically affected the use of the plough, which had persistent impacts on gender norms. We control for urban area fixed effects in all specifications, potentially mediating some of these concerns, but to investigate them more carefully, we used cultural data by ethnicity from the *Ethnographic Atlas* and Ashraf et al. (2020).

These data contain measures of the customs, practices, gender differences, community organizations, and traditional economies of each ethnic group. To construct controls for differences in culture, we merge these culture variables, which vary at the ethnicity level, to the 2000 census, recorded 12 years before our social capital outcomes are measured.[47] Note that while our controls for sorting, $\mathbf{X}_v$, do contain the shares of households that belong to different ethnic groups, those shares are measured in 2010 and some smaller groups have been aggregated, so there is independent variation in our culture controls that the $\mathbf{X}_v$ variables cannot explain.

The culture and social norms controls we use include the share of households the community who: (1) make a "bride price" payment at the time of marriage; (2) are from matrilocal societies; (3) are from patrilocal societies; (4) traditionally practiced male-led agriculture; (5) traditionally practiced female-lead agriculture; (6) practiced slavery historically; and (7) traditionally practiced polygamy. Appendix Table A.13 shows that including each of these controls separately (and together in the final column) does not change the results of our baseline trust in neighbors regressions (from Table 3). Column 4 of Appendix Table A.14 adds all of these controls together in the mean effects specifications. Overall, the estimated effects of density remain robust to including these cultural controls.

[47] Ideally, we would have used an older vintage of historical population census, but older village-level ethnicity shares are not available. For example, at best, the 1930 census contains ethnicity measures at the district level (Wang, 2021), and this variation would be washed out by our municipal area fixed effects.

Table A.13: The Effect of Density on Trust in Neighbors: Controlling for Culture and Social Norms

	Baseline		**Control Culture and Social Norms**							
	OLS	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**
Panel A: Only $\mathbf{X}_i$ and $\mathbf{C}_{2v}$ Controls	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
Log Density (2010)	-0.027***	-0.064***	-0.063***	-0.064***	-0.064***	-0.064***	-0.064***	-0.063***	-0.064***	-0.061***
	(0.005)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)
N	23,764	23,764	23,764	23,764	23,764	23,764	23,764	23,764	23,764	23,764
N Clusters	1,304	1,304	1,304	1,304	1,304	1,304	1,304	1,304	1,304	1,304
Adjusted R^2	0.035	0.034	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.036
Kleibergen-Paap Wald Rank F Stat		78.992	71.087	78.777	70.125	72.465	79.191	71.089	79.154	63.138
Panel B: Adding $\mathbf{X}_v$ Controls	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
Log Density (2010)	-0.008	-0.084***	-0.084***	-0.084***	-0.085***	-0.084***	-0.084***	-0.084***	-0.084***	-0.082***
	(0.009)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.027)
N	23,764	23,764	23,764	23,764	23,764	23,764	23,764	23,764	23,764	23,764
N Clusters	1,304	1,304	1,304	1,304	1,304	1,304	1,304	1,304	1,304	1,304
Adjusted R^2	0.041	0.041	0.041	0.041	0.041	0.041	0.042	0.041	0.041	0.042
Kleibergen-Paap Wald Rank F Stat		22.285	22.245	22.046	22.231	22.164	22.340	22.319	22.194	22.048
Social Norms Controls										
Bride Price	.	.	Yes	.	.	.	.	.	.	Yes
Matrilocal	.	.	.	Yes	.	.	.	.	.	Yes
Patrilocal	.	.	.	.	Yes	.	.	.	.	Yes
Male-lead Agriculture	.	.	.	.	.	Yes	.	.	.	Yes
Female-lead Agriculture	.	.	.	.	.	.	Yes	.	.	Yes
Former slavery	.	.	.	.	.	.	.	Yes	.	Yes
Polygomy	.	.	.	.	.	.	.	.	Yes	Yes
City FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each cell reports the coefficient on log population density in 2010 from equation (7) where the dependent variable is the 4-point index of trust in neighbors. Column 1 and 2 reproduce the OLS estimates and IV-lasso estimates from Table 3. Notice that the estimates are slightly different due to the imperfect merging to culture and social norms data. In Panel A, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. Panel B reports the full, unrestricted model. The specific variables we include in $\mathbf{X}_i$, $\mathbf{C}_{2v}$, and $\mathbf{X}_v$, as well as their coefficients, are reported in Appendix Table A.4 and Appendix Table A.5. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. From Column 3-9, we seperately includes different culture and social norms variable variable, as idicated from the bottom panel. In Column 10, we include all culture and social norms variables. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.14: Mean Effects of Density on Social Capital: Controls for Historical Development

	OLS	**IV-LASSO**	**IV-LASSO**	**IV-LASSO**
Panel A: Trust in Neighbors	**(1)**	**(2)**	**(3)**	**(4)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.075***	-0.118***	-0.138***	-0.110***
	(0.008)	(0.018)	(0.029)	(0.020)
2. Adding $\mathbf{X}_v$ Controls	-0.032**	-0.130***	-0.150***	-0.122***
	(0.014)	(0.040)	(0.044)	(0.040)
N Outcomes	2	2	2	2
N	44,020	44,020	44,020	44,020
N individuals	22,010	22,010	22,010	22,010
Panel B: Community Participation	**(1)**	**(2)**	**(3)**	**(4)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.025***	-0.045***	-0.053***	-0.043***
	(0.005)	(0.010)	(0.016)	(0.011)
2. Adding $\mathbf{X}_v$ Controls	-0.032***	-0.070***	-0.067***	-0.076***
	(0.007)	(0.024)	(0.026)	(0.023)
N Outcomes	5	5	5	5
N	106,992	106,992	106,992	106,992
N individuals	20,574	20,574	20,574	20,574
Panel C: Social Insurance	**(1)**	**(2)**	**(3)**	**(4)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.023***	-0.016	-0.006	-0.008
	(0.007)	(0.014)	(0.023)	(0.016)
2. Adding $\mathbf{X}_v$ Controls	0.017*	0.028	0.029	0.029
	(0.010)	(0.031)	(0.034)	(0.031)
N Outcomes	3	3	3	3
N	66,030	66,030	66,030	66,030
N individuals	22,010	22,010	22,010	22,010
Panel D: Intergroup Tolerance	**(1)**	**(2)**	**(3)**	**(4)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	0.054***	0.051**	0.045	0.040
	(0.010)	(0.023)	(0.034)	(0.025)
2. Adding $\mathbf{X}_v$ Controls	0.046***	0.013	0.037	0.007
	(0.017)	(0.050)	(0.054)	(0.051)
N Outcomes	2	2	2	2
N	39,156	39,156	39,156	39,156
N individuals	19,541	19,541	19,541	19,541
City FE	Yes	Yes	Yes	Yes
Historical Infrastructure Controls	.	.	Yes	.
Cultural and Social Norms Controls	.	.	.	Yes

Notes: Each cell reports the mean effect estimate, τ, of log population density in 2010 on groups of related outcomes, from equation (8). Outcome groupings are listed in the panel headers, and the outcomes themselves are reported in Table 1. All columns use post-double-selection IV-Lasso estimates, following Belloni et al. (2012). The first column reproduces our baseline IV-Lasso estimates from Table 4. For each panel, in row 1, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\boldsymbol{\Gamma}_1 = 0$, while row 2 reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Placebo Exercise. Finally, we conducted a placebo exercise to explore the role of our selected soil characteristics in communities where these IVs did not predict population density. The intuition behind this exercise is that in a subsample where we have no first stage relationship, there should also be no reduced-form relationship if the exclusion restriction is satisfied (Altonji et al., 2005; van Kippersluis and Rietveld, 2018).

We implemented this placebo exercise in two steps. First, we defined a subsample of communities where the first-stage relationship between our IVs and population density is weak by focusing on rural areas. We use the UN Statistical Commission's definition of rural areas as locations where population density is less than 300 inhabitants per square km (UN Statistical Commission, 2020).[48] Our placebo sample consists of 2,159 communities below this density threshold. Note that while we work with urban-area fixed effects in our main regressions, these have no analogue in the rural placebo communities. To make the comparison as close as possible, we use district fixed effects in the placebo specifications, but districts are often larger than urban areas, rendering our approach imperfect.

In Appendix Table A.15, we report reduced form coefficients on the soil characteristics IVs for all of the variables we use in our analysis. For these specifications, we include all individual and community controls, as well as controls for sorting. The table also reports p-values of a test of the joint significance of the soil characteristics for each dependent variable. Out of the 12 social capital outcomes, the selected soil characteristics are significantly related to only a single community participation variable. We take this, along with the other results in this section, as evidence largely in favor of the exclusion restriction.

[48] Although BPS has definitions for rural and urban areas in Indonesia, many communities that BPS classfies as rural are actually quite densely populated. Of the 6,751 communities in the 2012 Susenas, 4,208 communities were either classified as rural by the UN density threshold or by BPS. A total of 1,626 (38.6%) were categorized as rural by BPS but had density larger than 300 km. Only 201 communities (4.7%) were classified as rural based on population density but urban based on BPS definitions.

Table A.15: The Effects of Soil Characteristics on Social Capital in Rural Areas

	Trust in Neighbors		**Community Participation**				
	Trust Neighbors to Watch House	**Trust Neighbors to Tend Children**	**Join Community Groups**	**Join Religious Activities**	**Join Religious Activities Recently**	**Voluntary Public Good Provision**	**Join Community Activities Recently**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Soil bulk density at 60 cm depth (kg / m3)	-0.001	0.000	-0.001	-0.000	-0.000	-0.001	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Sand content at 60 cm depth (% (kg / kg))	-0.002	-0.002	0.008**	0.004	0.002	0.000	0.002
	(0.002)	(0.003)	(0.004)	(0.003)	(0.002)	(0.003)	(0.002)
Great Group 255: Dystropepts (Inceptisols) (0 1)	-0.016	-0.016	0.068	0.075*	0.030	-0.037	0.018
	(0.030)	(0.041)	(0.049)	(0.042)	(0.028)	(0.040)	(0.023)
Great Group 291: Haplustolls (Mollisols) (0 1)	0.075	0.152	0.041	-0.005	-0.093	-0.086	0.032
	(0.064)	(0.129)	(0.146)	(0.106)	(0.088)	(0.122)	(0.040)
Great Group 340: Haplorthox (Oxisols) (0 1)	-0.001	0.019	0.132***	0.089**	0.017	0.045	0.041*
	(0.028)	(0.038)	(0.044)	(0.036)	(0.023)	(0.035)	(0.023)
Great Group 402: Tropudults (Ultisols) (0 1)	-0.004	0.053	0.100**	0.078**	0.013	-0.002	0.030
	(0.024)	(0.035)	(0.039)	(0.035)	(0.022)	(0.033)	(0.023)
N	21,294	21,294	19,610	20,803	20,621	20,721	21,034
N Clusters	1,537	1,537	1,532	1,534	1,531	1,535	1,534
Adj. R^2	0.081	0.096	0.145	0.160	0.175	0.192	0.141
Adj. R^2 (Within)	0.007	0.011	0.045	0.050	0.046	0.078	0.048
$H_o : \beta = 0$ (p-value)	0.622	0.443	0.028	0.260	0.787	0.346	0.515
Kleibergen-Paap Wald Rank F Stat	4.327	4.327	4.569	4.387	4.115	4.418	4.391

	Social Insurance			**Intergroup Tolerance**	
	Ready to Help Neighbor	**Assist Unfortunate Neighbors**	**Easy Access to Neighbors' Help**	**Pleased with Non-coreligions**	**Pleased with Non-coethnics**
	(8)	**(9)**	**(10)**	**(11)**	**(12)**
Soil bulk density at 60 cm depth (kg / m3)	-0.000	-0.000	-0.001	0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Sand content at 60 cm depth (% (kg / kg))	0.001	0.000	-0.003	0.008**	-0.002
	(0.002)	(0.003)	(0.004)	(0.004)	(0.003)
Great Group 255: Dystropepts (Inceptisols) (0 1)	-0.030	0.006	-0.068	-0.032	0.012
	(0.030)	(0.038)	(0.048)	(0.055)	(0.042)
Great Group 291: Haplustolls (Mollisols) (0 1)	-0.107	-0.127	-0.072	-0.018	0.051
	(0.107)	(0.109)	(0.143)	(0.105)	(0.098)
Great Group 340: Haplorthox (Oxisols) (0 1)	0.006	0.041	-0.008	0.031	-0.020
	(0.028)	(0.033)	(0.043)	(0.043)	(0.034)
Great Group 402: Tropudults (Ultisols) (0 1)	-0.029	0.021	-0.009	0.025	-0.015
	(0.024)	(0.030)	(0.036)	(0.038)	(0.028)
N	21,294	21,294	21,294	18,817	19,514
N Clusters	1,537	1,537	1,537	1,398	1,439
Adj. R^2	0.092	0.110	0.089	0.231	0.151
Adj. R^2 (Within)	0.021	0.037	0.014	0.035	0.014
$H_o : \beta = 0$ (p-value)	0.452	0.741	0.697	0.319	0.843
Kleibergen-Paap Wald Rank F Stat	4.327	4.327	4.327	2.998	3.523

Notes: This table reports reduced-form regression coefficients of the dependent variable (listed in the column headers) on our selected soil characteristics in rural placebo areas (columns 3 and 4). All columns include city fixed effects and controls for $\mathbf{X}_i$, $\mathbf{C}_{2v}$, and $\mathbf{X}_v$. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. The "$H_o : \beta = 0$ (p-value)" row reports the p-value of an F-test for the null hypothesis that the coefficients on the soil characteristics variables are all equal to zero. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.16: The Effect of Density on Social Capital: Mean Effects, Adding More Controls

	Trust in Neighbors		**Community Participation**		**Social Insurance**		**Intergroup Tolerance**	
	OLS **(1)**	**IV-LASSO** **(2)**	**OLS** **(3)**	**IV-LASSO** **(4)**	**OLS** **(5)**	**IV-LASSO** **(6)**	**OLS** **(7)**	**IV-LASSO** **(8)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.068*** (0.008)	-0.112*** (0.017)	-0.021*** (0.005)	-0.041*** (0.010)	-0.020*** (0.006)	-0.016 (0.014)	0.055*** (0.010)	0.044** (0.021)
2. Adding $\mathbf{X}_v$ Controls	-0.026** (0.013)	-0.121*** (0.038)	-0.027*** (0.007)	-0.071*** (0.022)	0.014 (0.010)	0.017 (0.030)	0.045*** (0.016)	0.004 (0.045)
3. Adding More $\mathbf{W}_{2v}$ Controls	-0.022* (0.013)	-0.123*** (0.038)	-0.024*** (0.007)	-0.063*** (0.022)	0.012 (0.010)	0.020 (0.030)	0.044** (0.017)	-0.025 (0.046)
$H_o : \mathbf{\Gamma}_1 = 0$ (p-value)	0.041	0.052	0.000	0.000	0.623	0.595	0.210	0.190
$H_o : \tau_1 = \tau_2$ (p-value)	0.004	0.829	0.474	0.215	0.003	0.315	0.604	0.427
N Outcomes	2	2	5	5	3	3	2	2
N	47,784	47,784	116,144	116,144	71,676	71,676	42,517	42,517
N individuals	23,892	23,892	22,346	22,346	23,892	23,892	21,186	21,186
City FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each cell reports the mean effect estimate, τ, of log population density in 2010 on groups of related outcomes, from equation (8). Outcome groupings, and the outcomes themselves, are reported in Table 1. Column 1 reports OLS estimates, while Column 2 applies a post-double-selection lasso estimator to select the best soil characteristics instruments, following Belloni et al. (2012). In each panel, in row 1, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. Row 2 reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels. The additional $\mathbf{C}_{2v}$ controls include the community's distance to formal markets, if any restaurants exist, distance to schools, if there are any mobile phone or TV signals, the type of main water sources, if there are local community empowerment programs, the number of houses of worships, distance to medical facilities, and distance to maternal health facilities.

Table A.17: Summary Statistics: Social Capital Outcomes (IFLS 5)

Panel A: Trust in Neighbors	**Description**	**Code**	**Mean (sd)**	***N***
Trust neighbors to return wallet	Would you trust your neighbors to return your wallet if it was lost?	TR08	2.87 (1.05)	15,964
Trust neighbor to watch house	I would ask my neighbors to watch my house if I leave for a few days.	TR05	2.81 (0.57)	16,326
Trust neighbor to tend to children	I would leave children w/ neighbors for a few hours if I can't bring them along.	TR04	2.56 (0.64)	11,656
Panel B: Community Trust and Participation	**Description**	**Code**	**Mean (sd)**	***N***
Participate in village meetings?	In the last 12 months, did you participate in any village meetings?	PM16A	0.21 (0.41)	16,219
Participate in cooperatives?	In the last 12 months, did you participate in any cooperative?	PM16B	0.04 (0.19)	16,219
Participate in voluntary labor?	In the last 12 months, did you participate in any voluntary labor?	PM16C	0.27 (0.44)	16,219
Participate in program to improve the neighborhood?	In the last 12 months, did you participate in any program to improve the neighborhood?	PM16D	0.20 (0.40)	16,219
Participate in youth group activities?	In the last 12 months, did you participate in any youth group activities (Karang Taruna)?	PM16E	0.08 (0.27)	16,219
Participate in religious activities?	In the last 12 months, did you participate in any religious activities (Prayer groups, etc.)?	PM16F	0.52 (0.50)	16,219
Panel C: Social Insurance	**Description**	**Code**	**Mean (sd)**	***N***
Willing to help villagers in need	I am willing to help people in this village if they need it.	TR01	3.24 (0.46)	16,326
Panel D: Intergroup Tolerance	**Description**	**Code**	**Mean (sd)**	***N***
Trust own ethnic group more	Do you trust people with the same ethnicity as mine more than others	TR03	2.30 (0.66)	16,326
Trust own religious group more	Do you trust people with the same relgion as mine more than others	TR23	2.04 (0.67)	16,325
Tolerate diff. faith living in the same village	What if someone with a different faith lives in your village?	TR24	2.82 (0.59)	16,326
Tolerate diff. faith living in the same neighborhood	What if someone with a different faith lives in your neighborhood?	TR25	2.81 (0.58)	16,326
Tolerate diff. faith renting a room	What if someone with a different faith rents a room from you?	TR26	2.46 (0.71)	16,326
Tolerate diff. faith marrying relatives	What if someone with a different faith marries a close relative or child?	TR27	1.78 (0.72)	16,325
Tolerate diff. faith building house of worship nearby	What if people with a different faith build a house of worship in your community?	TR28	2.27 (0.77)	16,326

Notes: This table reports short titles, longer descriptions, and summary statistics for the social capital outcomes we analyze from the IFLS 5 data (2014-2015). Summary statistics were computed using data only from the sample of communities comprising metropolitan areas. The groupings of variables listed here correspond to the groupings used in the IFLS 5 mean effects analysis (e.g. Table 5).

Table A.18: Summary Statistics: Social Capital Outcomes (IFLS 4)

Panel A: Trust in Neighbors	**Description**	**Code**	**Mean (sd)**	***N***
Trust neighbors to return wallet	Would you trust your neighbors to return your wallet if it was lost?	TR08	2.99 (0.94)	13,784
Trust neighbor to watch house	I would ask my neighbors to watch my house if I leave for a few days.	TR05	2.84 (0.47)	14,092
Trust neighbor to tend to children	I would leave children w/ neighbors for a few hours if I can't bring them along.	TR04	2.61 (0.58)	10,459
Panel B: Community Trust and Participation	**Description**	**Code**	**Mean (sd)**	***N***
Participate in village meetings?	In the last 12 months, did you participate in any village meetings?	PM16A	0.21 (0.41)	14,085
Participate in cooperatives?	In the last 12 months, did you participate in any cooperative?	PM16B	0.03 (0.16)	14,085
Participate in voluntary labor?	In the last 12 months, did you participate in any voluntary labor?	PM16C	0.26 (0.44)	14,085
Participate in program to improve the neighborhood?	In the last 12 months, did you participate in any program to improve the neighborhood?	PM16D	0.18 (0.38)	14,085
Participate in youth group activities?	In the last 12 months, did you participate in any youth group activities (Karang Taruna)?	PM16E	0.06 (0.23)	14,085
Participate in religious activities?	In the last 12 months, did you participate in any religious activities (Prayer groups, etc.)?	PM16F	0.50 (0.50)	14,085
Panel C: Social Insurance	**Description**	**Code**	**Mean (sd)**	***N***
Willing to help villagers in need	I am willing to help people in this village if they need it.	TR01	3.14 (0.37)	14,094
Panel D: Intergroup Tolerance	**Description**	**Code**	**Mean (sd)**	***N***
Trust own ethnic group more	Do you trust people with the same ethnicity as mine more than others	TR03	2.41 (0.58)	14,093
Trust own religious group more	Do you trust people with the same relgion as mine more than others	TR23	2.26 (0.58)	14,094
Tolerate diff. faith living in the same village	What if someone with a different faith lives in your village?	TR24	2.86 (0.48)	14,094
Tolerate diff. faith living in the same neighborhood	What if someone with a different faith lives in your neighborhood?	TR25	2.83 (0.51)	14,094
Tolerate diff. faith renting a room	What if someone with a different faith rents a room from you?	TR26	2.50 (0.68)	14,093
Tolerate diff. faith marrying relatives	What if someone with a different faith marries a close relative or child?	TR27	1.75 (0.79)	14,093
Tolerate diff. faith building house of worship nearby	What if people with a different faith build a house of worship in your community?	TR28	2.33 (0.78)	14,094

Notes: This table reports short titles, longer descriptions, and summary statistics for the social capital outcomes we analyze from the IFLS 4 data (2007). Summary statistics were computed using data only from the sample of communities comprising metropolitan areas. The groupings of variables listed here correspond to the groupings used in the IFLS 4 mean effects analysis (e.g. Appendix Table A.19).

Table A.19: The Effect of Density on Social Capital: Mean Effects (IFLS 4)

	Trust in Neighbors		**Community Participation**		**Social Insurance**		**Intergroup Tolerance**	
	OLS	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
1. Only $\mathbf{X}_i$ and $\mathbf{W}_{2v}$ Controls	-0.047***	-0.070***	-0.027***	-0.025**	-0.025**	-0.038	0.088***	0.130***
	(0.008)	(0.016)	(0.006)	(0.010)	(0.013)	(0.024)	(0.007)	(0.013)
2. Adding $\mathbf{X}_v$ Controls	-0.010	-0.095**	-0.019***	-0.020	-0.023	-0.101*	0.043***	0.174***
	(0.011)	(0.042)	(0.007)	(0.029)	(0.016)	(0.060)	(0.007)	(0.033)
$H_o : \mathbf{\Gamma}_1 = 0$ (p-value)	0.666	0.695	0.000	0.000	0.697	0.635	0.000	0.000
$H_o : \tau_1 = \tau_2$ (p-value)	0.006	0.585	0.385	0.860	0.897	0.332	0.000	0.216
N Outcomes	3	3	6	6	1	1	7	7
N	39,418	39,418	86,748	86,748	14,465	14,465	101,253	101,253
N individuals	10,823	10,823	14,458	14,458	14,465	14,465	14,464	14,464
City FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each cell reports the mean effect estimate, τ, of log population density in 2010 on groups of related outcomes, from equation (8), but using the IFLS 5 data for outcomes. Outcome groupings are listed in the column headers, and the outcomes themselves are reported in Appendix Table A.18. Columns 1, 3, 5, and 7 report OLS estimates, while Columns 2, 4, 6, and 8 use post-double-selection IV-Lasso estimates, following Belloni et al. (2012). In row 1, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. Row 2 reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table A.20: Mean Effects, IFLS Panel Regressions (Single Index First Step)

	Trust in Neighbors		**Community Participation**		**Social Insurance**		**Intergroup Tolerance**	
	OLS	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**	**OLS**	**IV-LASSO**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Log density (2010)	-0.083*	-0.165***	-0.039	-0.046	-0.145*	-0.320*	0.091**	0.127*
	(0.044)	(0.061)	(0.039)	(0.065)	(0.083)	(0.164)	(0.041)	(0.075)
N	174	173	174	173	174	173	174	173
Adjusted R^2	0.011	0.001	-0.001	-0.008	0.011	-0.027	0.023	0.019
Kleibergen-Paap Wald Rank F Stat		36.885		36.885		36.885		36.885
Under Id. Test (KP Rank LM Stat)		42.291		42.291		42.291		42.291
p-Value		0.000		0.000		0.000		0.000
AR Wald Test (Weak IV Robust Inf.)		3.222		1.570		2.863		1.465
p-Value		0.014		0.185		0.025		0.215

Notes: This table reports panel regression results using IFLS 4/5 data. A single index is created by taking the average over all related outcome variables seperately for each group. Then we regress the 4 constructed single-index outcomes variables, namely trust in neighbors, community participation, social insurance, and intergroup tolerance, on log density. Outcome groupings are listed in the column headers. Columns 1, 3, 5, and 7 report OLS estimates, while Columns 2, 4, 6, and 8 use post-double-selection IV-Lasso estimates, following Belloni et al. (2012). All regressions are limited to the sample of IFLS villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

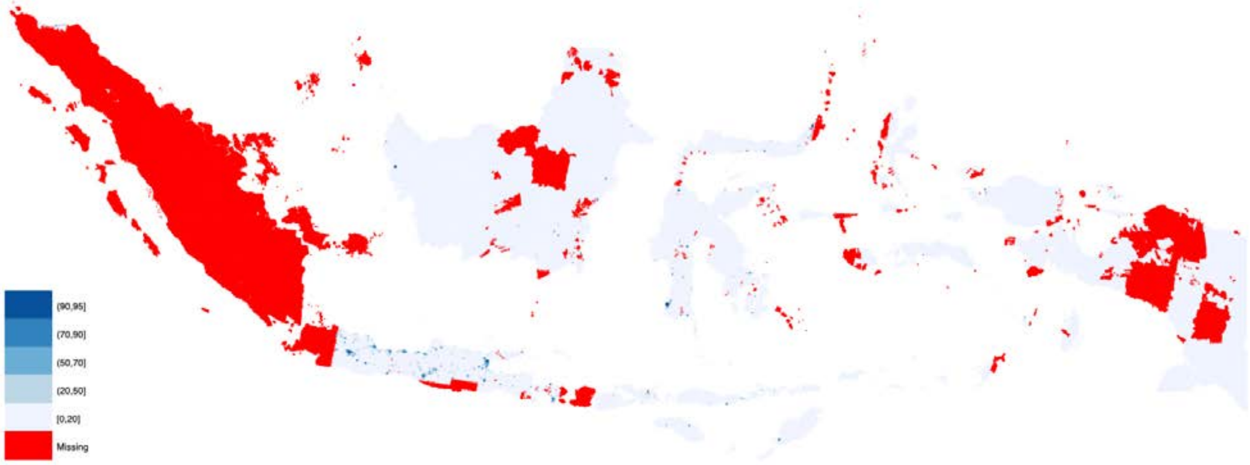
Table A.21: Summary Statistics: Mechanism Variables

Panel A: Individual level	**Description**	**Dataset**	**Mean (sd)**	**N**
Education	Years people spend in school. Individuals are classified as high (low) educated if school year is above (below) the sample median.	SUSENAS 2012	8.37 (4.60)	24,447
Private commuting mode to work (1 0)	If the commuting mode people choose to go to work is private or not. Private modes include private cars, private motorcycles and official cars. Non-private modes include non-motorized transportation and public transportation.	SUSENAS 2012	0.48 (0.50)	20,567
Income	Monthly net income (k Rupiah), in money and goods, earned by individuals from the main job. Individuals are classified into high and low income according to the median of the income distribution.	SUSENAS 2012	2056.93 (2968.66)	20,149

Panel B: City level	**Description**	**Dataset**	**Mean (sd)**	**N**
Commute Distance	Average commute distance in km. Cities are classified as high (low) commute distance if this index is above (below) the sample median	SAKERNAS 2009-2011	28.26 (21.80)	76
Property Crime	The city-level probability that a community within the city reports any incidence of property crime, weighted by population. Property crimes include theft and fraud. Cities are classified as suffering high (low) property crime risk if this index is above (below) the sample median.	PODES 2011	0.63 (0.17)	76
Violent Crime	The city-level probability that a community within the city reports any incidence of violent crime. Violent crimes include theft with violence, abuse, rape and homicide. Cities are classified as suffering high (low) violent crime risk if the index is above (below) the sample median.	PODES 2011	0.23 (0.14)	76

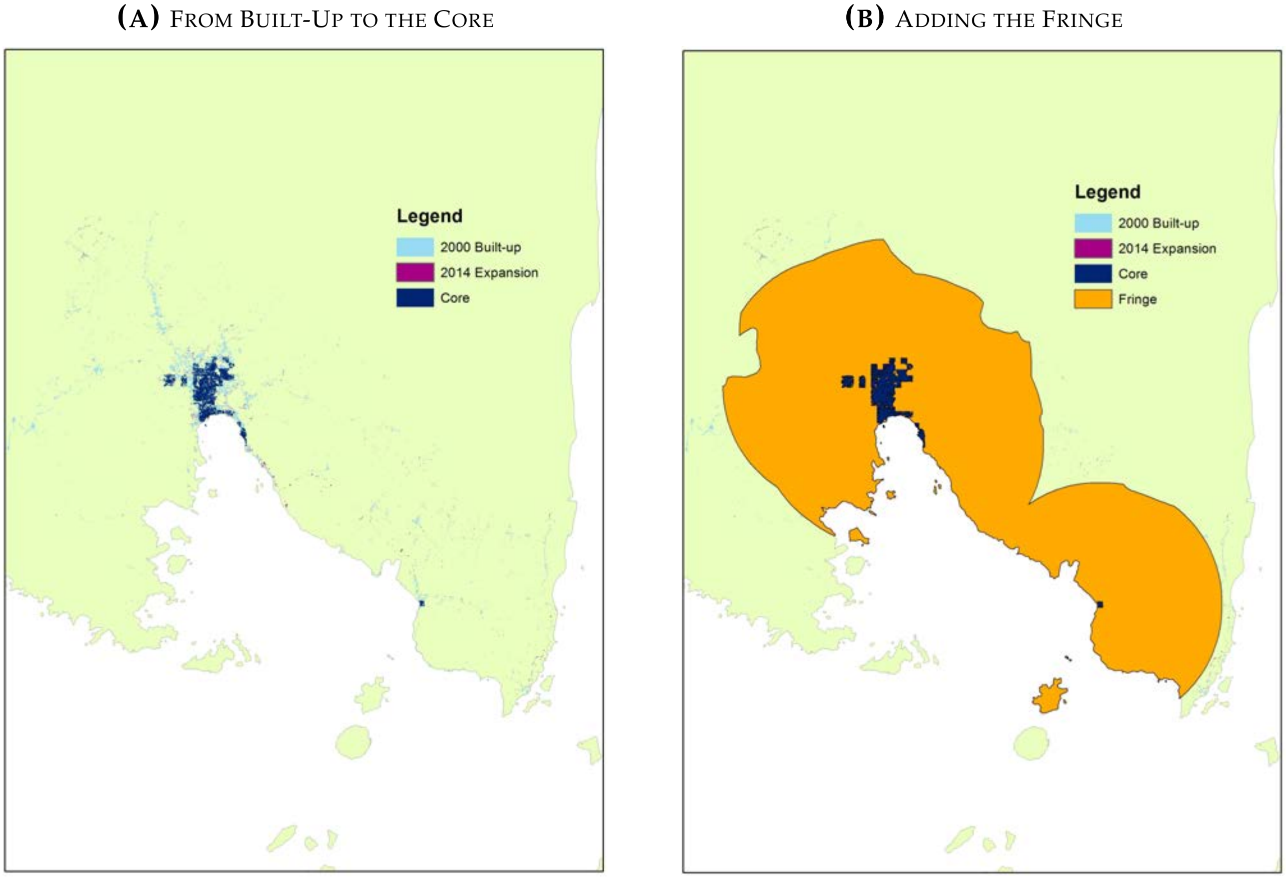
Notes: This table reports titles, descriptions for constructions, and summary statistics for the variables used in mechanism analysis for the *Susenas* data (2012). Summary statistics were computed using data only from the sample of communities comprising metropolitan areas.

Figure A.1: GHSL 1975: Built Up Area



Notes: This figure plots the built up extent of villages in Indonesia using the GHSL 1975 data. Locations with a larger percentage of built-up areas are shaded in darker blue. The red portions of this figure indicate areas where the 1975 data are missing. The figure assigns each village to the average of the GHSL 1975 raster for that village, using the 2010 village shapefile from BPS.

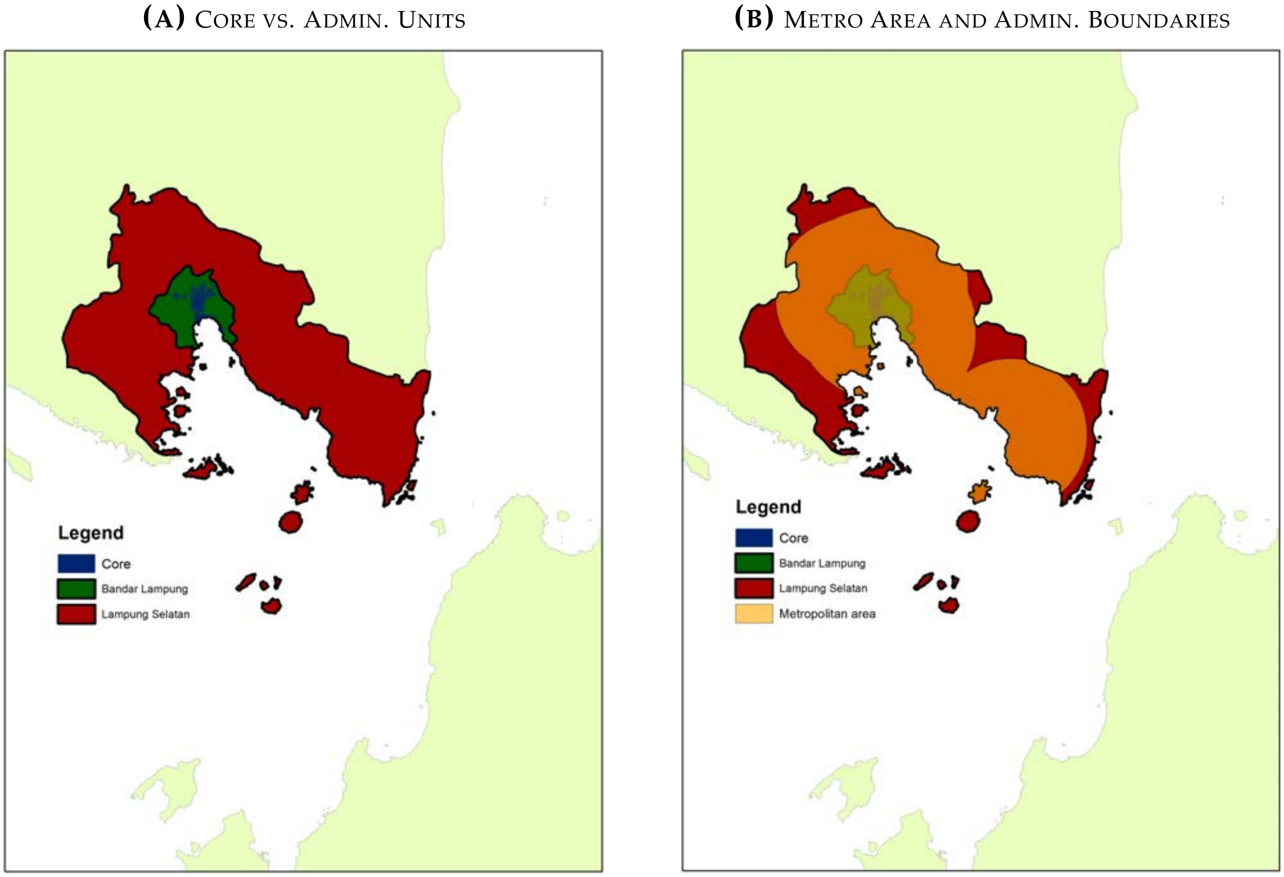
Figure A.2: Bandar Lampung Core and Fringe Identification


(A) From Built-Up to the Core

(B) Adding the Fringe

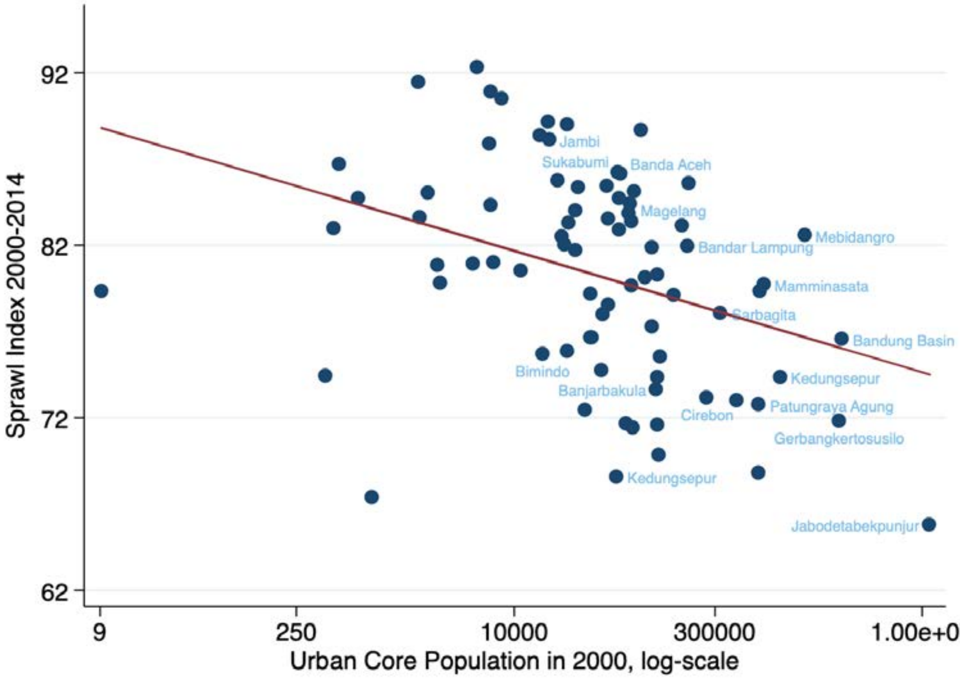
Notes: This figure illustrates the procedure we follow to identify the core and fringe of a metropolitan area from the built-up raster data. We use the metropolitan region corresponding to the city of Bandar Lampung as an example. Panel A reports the pixels in the map covered by built-up in 2000 (light blue) and those corresponding to new built-up by 2014 (dark violet). The light green polygon in the background depicts the Southern tip of Sumatra. The pixels of 2000 built-up areas identified as core by Burchfield et al. (2006) methodology are indicated in dark blue. For this city, a major core is visible at the center of the map, while a second smaller satellite core can be seen moving South-East along the shore. Panel B simply adds the fringe constructed around this core (in orange). The fringe is obtained as a 20km buffer around the core, which is then intersected with the boundaries of the administrative units belonging to this metropolitan area. Bandar Lampung comprises two administrative regions, as described in Appendix Figure A.3.

Figure A.3: Bandar Lampung Core-Fringe and Administrative Boundaries



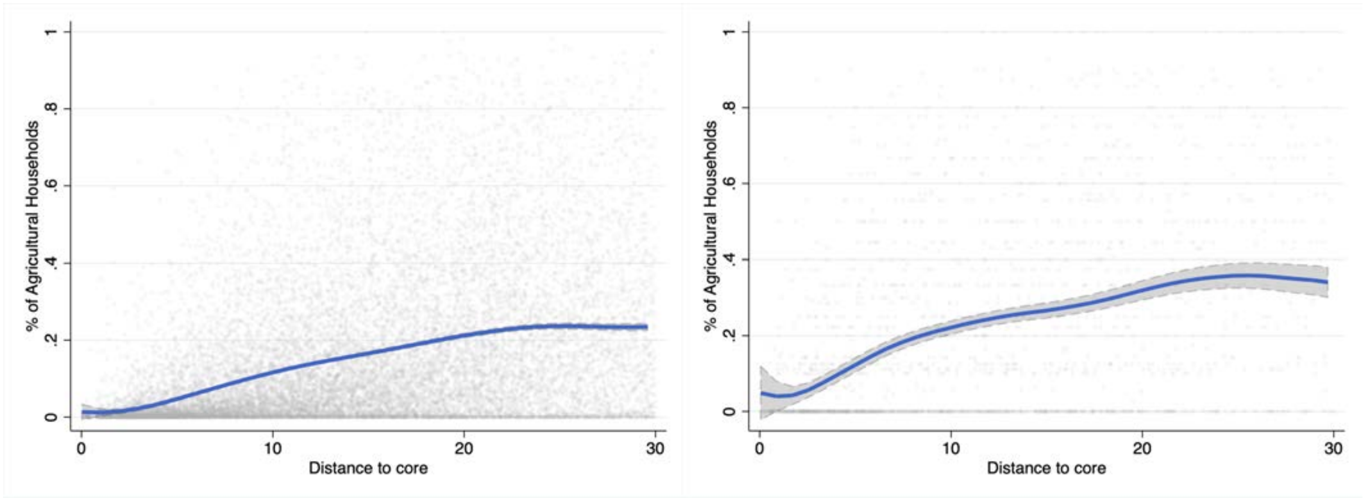
Notes: This figure is a continuation of Appendix Figure A.2 and illustrates how the metropolitan area identified by the Burchfield et al. (2006) methodology compares to the boundaries of the administrative units corresponding to it. The metropolitan area encompasses two administrative regions, which can be seen in Panel A. The dark green unit is the city (*kota*) of Bandar Lampung, while the crimson unit is the district (*kabupaten*) of Lampung Selatan. The core of the metropolitan area from Figure A.2 is also reported in dark blue. It is important to observe that the main portion of the core does not overlap with the administrative boundaries of the *kota*, being actually smaller, while the second part completely lies within the *kabupaten*. In Panel B, the identified metropolitan area is superimposed onto the administrative definition of it. The figure shows that the identified metropolitan area is also smaller than the simple union of the two administrative units. Moreover, it is evident that the area is delimited by the administrative boundaries when the radius of the fringe exceeds those boundaries.

Figure A.4: Urban Sprawl vs. Population of the Urban Core Area in 2000



Notes: This figure presents a scatterplot of the relationship between urban sprawl (from 2000-2014) and the population of the urban core area of each city in 2000. The estimated semi-elasticity of the linear regression line is -1.02 (p-value .00). Each point in the scatterplot represents a different city. The horizontal axis is expressed in log-scale.

Figure A.5: Agricultural HH Share and Distance to the CBD

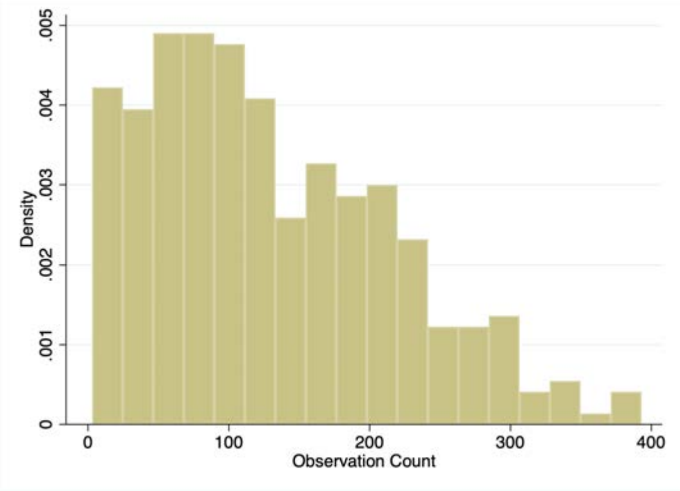


(A) Agricultural HH Share

(B) Agricultural HH Share, SUSENAS

Notes: These figures plot kernel-weighted local polynomial regressions of the share of agricultural households on distance to the CBD. An agricultural household is defined if all employed household members report working in agricultural sectors, based on 2010 census data. For communities in our sample, the average agricultural HH share is 16.3% (with a standard deviation 19%). These regressions use Epanechnikov kernels, rule-of-thumb bandwidths, and 3rd-degree polynomials.

Figure A.6: Number of Observations for Estimating α_{vt}



Notes: This figure reports a histogram of the number of individual-level observations used to estimate the α_{vt} terms from equation (9) in the first step of the Combes et al. (2008) procedure. The median number of observations is 114, but there is considerable variation across villages and years.

B Specification Checks

Appendix Tables B.1 and B.2 reports individual-outcome results from binary linear probability models, instead of the linear-index specifications used as the basis for Table 4. Before estimating these models, we coarsen the dependent variable into binary variables, where a 1 indicates positive social capital outcome and a 0 does not. Although the magnitudes of the estimated effects differ, the general conclusions of Table 4 are robust to this specification.

Table B.1: The Effect of Density on Trust in Neighbors and Community Participation (Linear Prob)

	OLS	IV-LASSO	Dep.Var	
	(1)	(2)	Mean (SE)	N
Panel A: Trust in Neighbors				
1. trust neighbor to watch house	-0.019***	-0.038***	0.842	23,892
	(0.004)	(0.008)	(0.002)	
... adding $\mathbf{X}_v$ controls	-0.007	-0.056***		
	(0.006)	(0.018)		
2. trust neighbor to tend children	-0.041***	-0.053***	0.639	23,892
	(0.006)	(0.012)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.022**	-0.051*		
	(0.009)	(0.027)		
Panel B: Community Participation				
3. join community group(s)	0.007	-0.004	0.475	22,346
	(0.005)	(0.010)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.007	-0.039*		
	(0.007)	(0.023)		
4. join religious activities	-0.019***	-0.016	0.643	23,498
	(0.005)	(0.010)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.020***	-0.012		
	(0.007)	(0.022)		
5. join religious activities recently	-0.015***	-0.034***	0.730	23,616
	(0.004)	(0.010)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.017**	-0.045**		
	(0.007)	(0.022)		
6. voluntary public good provision	-0.018***	-0.020*	0.523	23,081
	(0.005)	(0.011)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.020**	-0.029		
	(0.009)	(0.026)		
7. join community activities recently	-0.008*	-0.016*	0.798	23,603
	(0.005)	(0.009)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.011	-0.035*		
	(0.007)	(0.020)		
City FE	Yes	Yes		

Notes: Each cell reports the coefficient on log population density in 2010 from a separate linear probability model of (7) where the dependent variable is a binary coarsening of the outcome listed in the row header. Column 1 reports OLS estimates, while Column 2 applies a post-double-selection lasso estimator to select the best soil characteristics instruments, following Belloni et al. (2012). In the first row of each panel, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\boldsymbol{\Gamma}_1 = 0$. The second row of each panel reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table B.2: The Effect of Density on Social Insurance and Intergroup Tolerance (Linear Prob)

	OLS	**IV-LASSO**	**Dep.Var**	
	(1)	**(2)**	**Mean (SE)**	**N**
Panel C: Social Insurance				
8. ready to help neighbor	0.001	0.012	0.882	23,892
	(0.004)	(0.008)	(0.002)	
... adding $\mathbf{X}_v$ controls	0.007	0.026		
	(0.005)	(0.016)		
9. contribute to assist unfortunate neigbhors	-0.017***	-0.013	0.703	23,892
	(0.005)	(0.011)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.010	-0.003		
	(0.008)	(0.023)		
10. easily access to neighbors' help	-0.020***	-0.025**	0.625	23,892
	(0.006)	(0.012)	(0.003)	
... adding $\mathbf{X}_v$ controls	0.005	-0.016		
	(0.008)	(0.026)		
Panel D: Intergroup Tolerance				
11. pleased with non-coreligions	0.034***	0.039***	0.761	21,186
	(0.007)	(0.014)	(0.003)	
... adding $\mathbf{X}_v$ controls	0.027***	0.024		
	(0.010)	(0.028)		
12. pleased with non-coethnics	0.013**	0.004	0.826	21,331
	(0.005)	(0.012)	(0.003)	
... adding $\mathbf{X}_v$ controls	0.011	-0.022		
	(0.009)	(0.025)		
City FE	Yes	Yes		

Notes: Each cell reports the coefficient on log population density in 2010 from a separate linear regression of (7) where the dependent variable is the outcome listed in the row header. Column 1 reports OLS estimates, while Column 2 applies a post-double-selection lasso estimator to select the best soil characteristics instruments, following Belloni et al. (2012). In the first row of each panel, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. The second row of each panel reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

In Appendix Tables B.3 and B.4, we first coarsen the dependent variable into binary indicators. We then estimate the impact of density on individual outcomes using a binary probit model with instrumental variables. Our results on the impact of density on outcomes are robust to this limited dependent variable specification.

Table B.3: The Effect of Density on Trust in Neighbors and Community Participation (Binary Probit)

	OLS (1)	IV-LASSO (2)	Dep.Var Mean (SE)	N
Panel A: Trust in Neighbors				
1. trust neighbor to watch house	-0.080***	-0.146***	0.842	23,892
	(0.017)	(0.038)	(0.002)	
... adding $\mathbf{X}_v$ controls	-0.025	-0.203**		
	(0.027)	(0.085)		
2. trust neighbor to tend children	-0.115***	-0.155***	0.639	23,892
	(0.016)	(0.035)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.061**	-0.156*		
	(0.025)	(0.081)		
Panel B: Community Participation				
3. join community group(s)	0.020	-0.008	0.475	22,346
	(0.014)	(0.028)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.018	-0.106		
	(0.020)	(0.066)		
4. join religious activities	-0.052***	-0.045	0.643	23,498
	(0.014)	(0.028)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.057***	-0.033		
	(0.021)	(0.064)		
5. join religious activities recently	-0.049***	-0.107***	0.730	23,616
	(0.015)	(0.032)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.052**	-0.149**		
	(0.023)	(0.074)		
6. voluntary public good provision	-0.048***	-0.051	0.523	23,081
	(0.014)	(0.031)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.052**	-0.075		
	(0.023)	(0.072)		
7. join community activities recently	-0.031*	-0.056	0.798	23,603
	(0.017)	(0.035)	(0.003)	
... adding $\mathbf{X}_v$ controls	-0.044	-0.136*		
	(0.027)	(0.081)		
City FE	Yes	Yes		

Notes: Each cell reports the coefficient on log population density in 2010 from a separate binary probit model of (7). All columns report maximum likelihood estimates; column 1 reports them without using instruments, while column 2 use the instruments listed in the column headers. In column 2, we follow Belloni et al. (2012) and use a lasso procedure to select the best soil characteristics instruments before implementing the ML estimator. In the first row of each panel, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. The second row of each panel reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table B.4: The Effect of Density on Social Insurance and Intergroup Tolerance (Binary Probit)

	OLS (1)	IV-LASSO (2)	Dep.Var Mean (SE)	N
Panel C: Social Insurance				
8. ready to help neighbor	0.004 (0.018)	0.049 (0.039)	0.882 (0.002)	23,892
... adding $\mathbf{X}_v$ controls	0.035 (0.026)	0.121 (0.085)		
9. contribute to assist unfortunate neigbhors	-0.050*** (0.015)	-0.040 (0.032)	0.703 (0.003)	23,892
... adding $\mathbf{X}_v$ controls	-0.031 (0.023)	-0.006 (0.071)		
10. easily access to neighbors' help	-0.054*** (0.015)	-0.065* (0.034)	0.625 (0.003)	23,892
... adding $\mathbf{X}_v$ controls	0.016 (0.023)	-0.036 (0.073)		
Panel D: Intergroup Tolerance				
11. pleased with non-coreligions	0.124*** (0.023)	0.146*** (0.049)	0.761 (0.003)	21,186
... adding $\mathbf{X}_v$ controls	0.101*** (0.037)	0.086 (0.107)		
12. pleased with non-coethnics	0.057** (0.023)	0.032 (0.051)	0.826 (0.003)	21,331
... adding $\mathbf{X}_v$ controls	0.050 (0.037)	-0.067 (0.122)		
City FE	Yes	Yes		

Notes: Each cell reports the coefficient on log population density in 2010 from a separate binary probit model of (7). All columns report maximum likelihood estimates; column 1 reports them without using instruments, while column 2 use the instruments listed in the column headers. In column 2, we follow Belloni et al. (2012) and use a lasso procedure to select the best soil characteristics instruments before implementing the ML estimator. In the first row of each panel, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. The second row of each panel reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

In Appendix Tables B.5 and B.6, we estimate effects using ordered probit models with instruments, adopting the control function procedure proposed by Chesher and Rosen (2019). Our results on the impact of density on outcomes are robust to this limited dependent variable specification.

Table B.5: The Effect of Density on Trust in Neighbors and Community Participation (Ordered Probit Using Control Function)

	OLS (1)	IV-LASSO (2)	Dep.Var Mean (SE)	N
Panel A: Trust in Neighbors				
1. trust neighbor to watch house	-0.063*** (0.013)	-0.126*** (0.029)	2.916 (0.003)	23,892
... adding $\mathbf{X}_v$ controls	-0.014 (0.021)	-0.167*** (0.063)		
2. trust neighbor to tend children	-0.100*** (0.014)	-0.141*** (0.030)	2.648 (0.004)	23,892
... adding $\mathbf{X}_v$ controls	-0.042** (0.021)	-0.134** (0.066)		
Panel B: Community Participation				
3. join community group(s)	0.019* (0.012)	0.001 (0.024)	2.365 (0.006)	22,346
... adding $\mathbf{X}_v$ controls	-0.015 (0.017)	-0.090* (0.054)		
4. join religious activities	-0.045*** (0.011)	-0.050** (0.023)	2.689 (0.005)	23,498
... adding $\mathbf{X}_v$ controls	-0.043** (0.017)	-0.057 (0.052)		
6. voluntary public good provision	-0.030** (0.012)	-0.041 (0.027)	2.507 (0.005)	23,081
... adding $\mathbf{X}_v$ controls	-0.023 (0.019)	-0.051 (0.060)		
City FE	Yes	Yes		

Notes: Each cell reports the coefficient on log population density in 2010 from a separate ordered probit model of (7). Column 1 reports maximum likelihood (ML) estimates, while column 2 adopts the control function (CF) procedure described by Chesher and Rosen (2019), together with the instruments listed in the column headers. In column 2, we follow Belloni et al. (2012) and use a lasso procedure to select the best soil characteristics instruments before implementing the control function estimator. In the first row of each panel, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. The second row of each panel reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

Table B.6: The Effect of Density on Social Insurance and Intergroup Tolerance (Ordered Probit Using Control Function)

	OLS (1)	IV-LASSO (2)	Dep.Var Mean (SE)	N
Panel C: Social Insurance				
8. ready to help neighbor	0.005 (0.014)	0.038 (0.029)	2.981 (0.003)	23,892
... adding $\mathbf{X}_v$ controls	0.038* (0.021)	0.081 (0.062)		
9. contribute to assist unfortunate neigbhors	-0.031** (0.012)	-0.026 (0.026)	2.809 (0.005)	23,892
... adding $\mathbf{X}_v$ controls	-0.009 (0.018)	0.015 (0.057)		
10. easily access to neighbors' help	-0.043*** (0.013)	-0.064** (0.028)	2.653 (0.005)	23,892
... adding $\mathbf{X}_v$ controls	0.025 (0.020)	-0.018 (0.061)		
Panel D: Intergroup Tolerance				
11. pleased with non-coreligions	0.102*** (0.019)	0.099** (0.041)	2.736 (0.004)	21,186
... adding $\mathbf{X}_v$ controls	0.088*** (0.030)	0.018 (0.085)		
12. pleased with non-coethnics	0.041** (0.019)	0.009 (0.039)	2.822 (0.004)	21,331
... adding $\mathbf{X}_v$ controls	0.039 (0.030)	-0.101 (0.086)		
City FE	Yes	Yes		

Notes: Each cell reports the coefficient on log population density in 2010 from a separate ordered probit model of (7). Column 1 reports maximum likelihood (ML) estimates, while columns 2-4 adopt the control function (CF) procedure described by Chesher and Rosen (2019), together with the instruments listed in the column headers. In column 2, we follow Belloni et al. (2012) and use a lasso procedure to select the best soil characteristics instruments before implementing the control function estimator. In the first row of each panel, we only control for $\mathbf{X}_i$ and $\mathbf{C}_{2v}$, setting $\mathbf{\Gamma}_1 = 0$. The second row of each panel reports the full, unrestricted model. All regressions are limited to the sample of villages within urban areas and include city-fixed effects. Robust standard errors, clustered at the subdistrict-level, are reported in parentheses. */**/*** denotes significant at the 10% / 5% / 1% levels.

C Empirical Strategy Appendix

In this appendix, we describe how we address the two key identification challenges that confound estimates of the relationship between density and social capital: (1) sorting of individuals with lower or higher costs to contributing to social capital and (2) the simultaneous determination of density and social capital by unobservable place-specific variables. Our empirical strategy builds on the control function approach of Altonji and Mansfield (2018) for bounding the variance of group-level treatment effects in the presence of sorting into groups, but we add instruments to point identify group treatment effects.

C.1 Sorting into Communities

Let i index households and let $v \in \{1, ..., V\}$ index the discrete set of communities comprising different metropolitan areas in Indonesia. Household i's consumer surplus from choosing to live in community v is given by the following expression:

$$U_i(v) = \mathbf{W}_i \mathbf{A}_v - P_v + \varepsilon_{iv} \, , \tag{12}$$

where $\mathbf{A}_v$ represents a $(K \times 1)$ vector of amenities that characterize community v, P_v is the price of living in community v, and ε_{iv} is an idiosyncratic component specific to individual i's tastes for living in community v. The term $\mathbf{W}_i$ represents a $(1 \times K)$ vector of weights measuring household i's willingness to pay for different components of the amenity vector.

We partition $\mathbf{W}_i$ into three components: (1) $\mathbf{X}_i$, a vector of individual-level observables that influence tastes for amenities and social capital outcomes; (2) $\mathbf{X}_i^U$, a vector of individual-level unobservables that influence tastes for amenities and social capital outcomes; and (3) $\mathbf{Q}_i$, a vector of variables (both observed and unobserved) that may influence preferences over amenities and sorting but have no impact on social capital outcomes:

$$\mathbf{W}_i = \mathbf{X}_i \mathbf{\Theta} + \mathbf{X}_i^U \mathbf{\Theta}^U + \mathbf{Q}_i \mathbf{\Theta}^Q \, ,$$

where $\mathbf{\Theta}$, $\mathbf{\Theta}^U$, and $\mathbf{\Theta}^Q$ are the respective willingness to pay coefficients. Note that we define $\mathbf{X}_i$ and $\mathbf{X}_i^U$ so that they represent the complete set of individual factors that determine social capital outcomes. As emphasized by Altonji and Mansfield (2018), this formulation allows for a fairly general pattern of relationships between different individual characteristics (both observable and unobservable) and tastes for community characteristics, subject to the additive separability of the indirect utility function, equation (1).

We assume that households take prices, P_v, and amenities, $\mathbf{A}_v$, as given when making their location decisions, and that households choose the community that maximizes (1) using all information available to them. This information set includes housing prices in different locations, the vectors of amenities in those locations, the full set of preference weights, $\mathbf{W}_i$, and realizations of the idiosyncratic component, ε_{iv} for all $v \in \{1, ..., V\}$. Let $v(i)$ denote the optimal community choice for household i.

Altonji and Mansfield (2018) prove that given this setup and under a relatively weak set of additional assumptions, the community-level expectation of individual-level unobservables that influence the social capital outcome, denoted by $\mathbf{X}_v^U \equiv \mathbb{E}[X_i^U \,|\, v(i) = v\,]$, is linearly dependent on group-level average observables, $\mathbf{X}_v \equiv \mathbb{E}[X_i \,|\, v(i) = v\,]$. The proposition is restated here in full:

Proposition C.1. *(Altonji and Mansfield, 2018): Assume the following assumptions hold:*

- ***Assumption A1**: Preferences are given by equation (12).*
- ***Assumption A2**: Households take prices, P_v, and amenities, $\mathbf{A}_v$ as given when choosing locations, and they face a common choice set.*
- ***Assumption A3**: The idiosyncratic preference components, ε_{iv}, are mean zero and are independent of $\mathbf{X}_i$, $\mathbf{X}_i^U$, $\mathbf{Q}_i$, and $\mathbf{A}_v$ for all v.*

- ***Assumption A4***: *$\mathbb{E}[\mathbf{X}_i \mid \mathbf{W}_i\,]$ and $\mathbb{E}[\mathbf{X}_i^U \mid \mathbf{W}_i\,]$ are linear in $\mathbf{W}_i$.*
- ***Assumption A5***: (spanning assumption)*: Let $\mathbf{\Pi}_{\mathbf{X}^U\mathbf{X}}$ denote the matrix of partial regression coefficients relating the vector $\mathbf{X}_i^U$ to the vector $\mathbf{X}_i$. The row space of the WTP coefficient matrix $\widetilde{\mathbf{\Theta}} \equiv \left[\mathbf{\Theta} + \mathbf{\Pi}_{\mathbf{X}^U\mathbf{X}}\mathbf{\Theta}^U\right]$ spans the row space of the WTP coefficient matrix $\mathbf{\Theta}^U$ relating tastes for $\mathbf{A}$ to $\mathbf{X}_i^U$. That is,*

  $$\mathbf{\Theta}^U = \mathbf{R}\widetilde{\mathbf{\Theta}}$$

  for some $(L^U \times L)$ matrix $\mathbf{R}$.

Then, the expectation $\mathbf{X}_v^U \equiv \mathbb{E}[\mathbf{X}_i^U \mid v(i) = v\,]$, is linearly dependent on group-level average observables, $\mathbf{X}_v \equiv \mathbb{E}[\mathbf{X}_i \mid v(i) = v\,]$.

The proof of this proposition is in Altonji and Mansfield (2018). The intuition behind the argument is that sorting creates two vector-valued mappings: (1) a mapping between group level averages of observables in community v and the amenities in that community, $\mathbf{X}_v = \mathbf{f}\left(\mathbf{A}_v\right)$; and (2) a mapping between group-level averages of unobservables in community v and amenities, $\mathbf{X}_v^U = \mathbf{f}^U\left(\mathbf{A}_v\right)$. The authors provide conditions under which the first mapping, $\mathbf{f}$, is invertible, so we can write: $\mathbf{X}_v^U = \mathbf{f}^U\left(\mathbf{f}^{-1}\left(\mathbf{X}_v\right)\right)$. Under an additional assumption, the relationship between $\mathbf{X}_v^U$ and $\mathbf{X}_v$ induced by inverting these vector-valued functions is actually linear.

Assumptions A1-A3 of Proposition C.1 are discussed in detail in Altonji and Mansfield (2018), but we will make a few comments about Assumptions A4 and A5 here. A sufficient condition for Assumption A4 is that the joint distribution of $[\mathbf{X}_i, \mathbf{X}_i^U, \mathbf{Q}_i]$ is a member of the continuous elliptical class (Agrawal et al., 2019). Because our application uses several discrete variables in $\mathbf{X}_i$, this sufficient condition will not be satisfied. Altonji and Mansfield (2018) explain that this will introduces an approximation error in the control function, but they also provide reasons to believe that this approximation error will be small in practice.

The strongest assumption of Proposition C.1 is the spanning assumption, Assumption A5. Another way of stating this assumption is that the coefficient vectors $\mathbf{\Theta}^U$, which relate tastes for amenities to elements of $\mathbf{X}_i^U$, need to be linear combinations of $\mathbf{\Theta}$, which relate tastes for amenities to elements of $\mathbf{X}_i$ and/or elements of $\mathbf{X}_i^U$ that are correlated with $\mathbf{X}_i$. One of the two sufficient conditions for Assumption A5 to hold is that $\mathbf{f}$ is invertible. A necessary condition for invertibility is that the dimension of $\mathbf{A}^{\mathbf{X}}$, the subset of amenities that affect the distribution of community averages, is less than the number of elements in $\mathbf{X}_v$. This would occur if $\mathbb{V}\left(\mathbf{X}_v\right)$ is rank deficient.

In our empirical implementation, we use a vector of 38 variables constructed from unit-level 2010 census data to measure $\mathbf{X}_v$. These variables include the community's average age, years of schooling, household size, the percentage of the community that is female, the percent who self-identify with different religions and with ethnicities, the share of different types of employment status and marital status, and the share who speak Indonesian at home.[49] Appendix Table A.1 reports a principal components analysis of these 38 $\mathbf{X}_v$ variables. In our urban Susenas sample (column 2), only 27 factors explain 95 percent of the total variation in $\mathbf{X}_v$, 32 factors explain 99 percent of the total variation in $\mathbf{X}_v$, and 37 factors explain 100 percent of the total variation in $\mathbf{X}_v$. This suggests that for the urban Susenas sample, $\mathbf{X}_v$ is rank deficient.

Appendix Table A.2 also formally tests hypotheses about the rank of the $\mathbf{X}_v$ covariance matrix, using a test proposed by Kleibergen and Paap (2006). We find that for the full Susenas sample, we cannot reject the null hypothesis that the rank of the variance-covariance matrix of $\mathbf{X}_v$ is 34 against the alternative that it is 35 or greater. For the urban sample, we cannot reject the null hypothesis that the rank of the variance-covariance matrix of $\mathbf{X}_v$ is 28 against the alternative that it is 29 or greater. The results from both Appendix Table A.1 and Appendix Table A.2 suggest that because $\mathbf{X}_v$ is rank deficient, $\mathbf{f}$ is likely invertible, so that $\mathbf{X}_v$ can be used as a control function for sorting on unobservables.

[49] If the 2010 census data are inaccurate, they could provide potentially noisy measures of $\mathbf{X}_v$, the expected values of observable characteristics in community v. However, Altonji and Mansfield (2018) provide a Monte Carlo analysis suggesting that even with small samples from survey data (i.e., $N = 20$), we can approximate $\mathbf{X}_v$ fairly well.

C.2 Production of Social Capital

After households choose locations, we assume that a social capital outcome for household i living in community v, denoted by y_{vi}, is produced according to the following linear, additively separable function:

$$y_{vi} = \mathbf{X}_i \beta + x_i^U + \theta \log \text{density}_v + \mathbf{C}_v \mathbf{\Gamma} + c_v^U + \eta_{vi} + \xi_{vi} \,. \tag{13}$$

Because many outcomes recorded in the 2012 Susenas data are either binary or take on discrete values (often 4 point scales), y_{vi} is the continuous latent variable that determines these values. Equation (2) is composed of three sets of terms: (1) an individual component; (2) a community-level component; and (3) an idiosyncratic component. We describe each of these components in detail.

The individual component, $\mathbf{X}_i \beta + x_i^U$, includes a row vector, $\mathbf{X}_i$, collecting individual i's observed attributes that affect average willingness to contribute to the social capital outcome. The parameter β measures how those observed attributes affect y_{vi}. The second part of the individual component consists of a scalar, $x_i^U \equiv \mathbf{X}_i^U \beta^{\mathbf{U}}$, which summarizes the contribution of unobserved individual characteristics ($\mathbf{X}_i^U$) to social capital outcomes.

The community-level component, $\theta \log \text{density}_v + \mathbf{c}_v \mathbf{\Gamma} + c_v^U$, contains three terms. The first is a measure of log population density at the community level. The key object of interest, θ, is the parameter that measures the semi-elasticity of social capital outcomes with respect to density. The second component is a row vector, $\mathbf{C}_v$, capturing the influence of other observed community-level characteristics on social capital outcomes. Finally, the third term, $c_v^U \equiv \mathbf{C}_v^U \mathbf{\Gamma}^U$, represents a scalar that summarizes the contribution of unobserved neighborhood characteristics to the social capital outcome.

Finally, the idiosyncratic component, $\eta_{vi} + \xi_{vi}$, also contains two terms. The first term, η_{vi}, captures unobserved variation in community contributions to social capital among individuals who live in that community. Some factors correlated with η_{vi} may be captured by observed and unobserved community-level variables (e.g. $\log \text{density}_v$, $\mathbf{C}_v$ and c_v^U). The second term, ξ_{vi}, captures other influences to household i's social capital outcome that are determined after that household arrives in community v, but are unpredictable given $\mathbf{X}_i$, x_i^U, $\log \text{density}_v$, $\mathbf{C}_v$, c_v^U, and η_{vi}. Such influences could include local labor market shocks that make it harder or easier to participate in the community, or shocks to local public goods that influence different individuals in certain areas.

We partition the group-level observed variables (excluding log density) into $\mathbf{C}_v = [\mathbf{X}_v, \mathbf{C}_{2v}]$, and we partition their coefficients analogously, so that $\mathbf{\Gamma} = [\mathbf{\Gamma}_1, \mathbf{\Gamma}_2]$. The term $\mathbf{X}_v$ includes community averages of individual-level observables, while the term $\mathbf{C}_{2v}$ includes community-level characteristics that are not mechanically related to community composition. In our baseline specifications, these include pre-determined, exogenous natural amenities, such as elevation, ruggedness, and distance to the coast or rivers, which may make it easier or harder to sustain a social capital outcome. This notation lets us we rewrite equation (13) as follows:

$$y_{vi} = \mathbf{X}_i \beta + x_i^U + \theta \log \text{density}_v + \mathbf{X}_v \mathbf{\Gamma}_1 + \mathbf{C}_{2v} \mathbf{\Gamma}_2 + c_v^U + \eta_{vi} + \xi_{vi} \,. \tag{14}$$

Note that because of the assumptions described above, adding $\mathbf{X}_v$ effectively controls both for sorting on observables and on unobservables in the production of social capital. Although a typical control function procedure would use a non-linear or semi-parametric control function, the spanning assumption (Assumption A5) implies that this function is linear, and we need only to add these controls linearly.

In Altonji and Mansfield (2018), the authors use controls for sorting on observables and unobservables to bound the contribution of group treatment effects (e.g., schools or neighborhoods) to the total variance in outcomes. When estimating this overall group treatment effect, controlling for group averages eliminates the sorting bias, but it also controls for too much, because peer effects may depend on these group averages.[50] Consequently, they can only obtain a lower bound on the overall importance of school or neighborhood effects on outcomes.

[50] More subtly, group averages will also absorb part of the unobserved group quality component that is both orthogonal to observed group characteristics and correlated with amenities that families consider when choosing where to live.

In what follows, we describe how to extend their approach with an instrument for a particular group attribute (namely density) to point identify the effect of that attribute on outcomes in a way that is unconfounded by sorting.[51]

C.3 An Instrumental Variables (IV) Estimator for θ

Recall that $\widetilde{\mathbf{X}}_{iv} \equiv [\mathbf{X}_i, \mathbf{X}_v, \mathbf{C}_{2v}]$ collects observed variables that do not include density. A two-stage least squares (IV) estimator for $(\theta, \widetilde{\beta})$ solves the following two moment equations:

$$\begin{aligned} 0 &= \mathbf{Z}'(\mathbf{y} - \widehat{\theta}_{\text{2SLS}} \log \text{density} - \widetilde{\mathbf{X}}\widehat{\widetilde{\beta}}_{\text{2SLS}}) \\ 0 &= \widetilde{\mathbf{X}}'(\mathbf{y} - \widehat{\theta}_{\text{IV}} \log \text{density} - \widetilde{\mathbf{X}}\widehat{\widetilde{\beta}}_{\text{IV}}). \end{aligned}$$

The second equation can be used to solve for $\widehat{\widetilde{\beta}}_{IV}$ as follows:

$$\widehat{\widetilde{\beta}}_{\text{IV}} = \left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}' \left(\mathbf{y} - \widehat{\theta}_{\text{IV}} \log \text{density}\right).$$

Plugging this expression into the first equation gives us the following expression for the IV estimator of θ, our parameter of interest:

$$\widehat{\theta}_{\text{IV}} = \left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}} \log \text{density}\right)^{-1} \mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\mathbf{y}$$

where $\mathbf{M}_{\widetilde{\mathbf{X}}} = \left(\mathbf{I} - \widetilde{\mathbf{X}}\left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\right)$ is the standard orthogonal projection matrix for $\widetilde{\mathbf{X}}$.

C.4 Bias of $\widehat{\theta}_{\text{IV}}$

An expression for the bias of $\widehat{\theta}_{\text{IV}}$ is the following:

$$\begin{aligned} \text{Bias}\left(\widehat{\theta}_{\text{IV}}\right) &= \mathbb{E}\left[\widehat{\theta}_{IV}\right] - \theta \\ &= \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}} \log \text{density}\right)^{-1} \mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\mathbf{y}\right] - \theta \\ &= \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}} \log \text{density}\right)^{-1} \mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\ \mathbb{E}\left[\mathbf{y} \,\middle|\, \mathbf{Z}, \widetilde{\mathbf{X}}, \log \text{density}\right]\right] - \theta \\ &= \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}} \log \text{density}\right)^{-1} \mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\ \left(\theta \log \text{density}_v + \widetilde{\mathbf{X}}_{iv}\widetilde{\beta} + \mathbb{E}\left[u_{iv} \,\middle|\, \mathbf{Z}, \widetilde{\mathbf{X}}, \log \text{density}\right]\right)\right] - \theta \\ &= \theta + 0 + \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}} \log \text{density}\right)^{-1} \mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\ \mathbb{E}\left[u_{iv} \,\middle|\, \mathbf{Z}, \widetilde{\mathbf{X}}, \log \text{density}\right]\right] - \theta \\ &= \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}} \log \text{density}\right)^{-1} \mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\ \mathbb{E}\left[u_{iv} \,\middle|\, \mathbf{Z}, \widetilde{\mathbf{X}}, \log \text{density}\right]\right] \\ &= \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}} \log \text{density}\right)^{-1} \mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\ \mathbb{E}\left[\underbrace{x_i^U}_{\text{(A)}} + \underbrace{c_v^U}_{\text{(B)}} + \underbrace{\eta_{vi}}_{\text{(C)}} + \underbrace{\xi_{vi}}_{\text{(D)}} \,\middle|\, \mathbf{Z}, \widetilde{\mathbf{X}}, \log \text{density}\right]\right] \end{aligned}$$

This bias expression contains four terms, which we describe separately. For term (A), the sorting model described by Altonji and Mansfield (2018), together with their assumptions A1-A5, delivers their proposition 1, which is

[51] This insight was actually discussed in the original Altonji and Mansfield (2018) paper. From the introduction, p. 2094, with emphasis added: "... [T]he fact that controlling for the group averages eliminates bias from sorting implies that the causal effects ($\mathbf{\Gamma}$) of *particular school inputs or policies* (in $\mathbf{Z}_s$) can be **point identified** in situations where bias from omitted neighborhood/school characteristics in z_s^U is not a problem or can be addressed through a *complementary instrumental variables* scheme."

namely that the expectation of x_v^U is linearly dependent on $\mathbf{X}_v$, the group-level observables. This means that we have the following:

$$\text{(A) :} \qquad \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\text{log density}\right)^{-1}\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\mathbb{E}\left[x_i^U\middle|\,\mathbf{Z},\widetilde{\mathbf{X}},\text{log density}\right]\right]=0$$

since the expectation of x_i^U is in the space spanned by the columns of $\widetilde{\mathbf{X}}$.

For term (C), note that $\mathbf{C}_{2v}$ is defined as a vector of village-level characteristics not mechanically related to sorting, so $\mathbf{C}_{2v}$ is uncorrelated with η_{vi} by definition. Hence, we can write:

$$\eta_{vi}=\mathbf{X}_i\mathbf{P}_{\mathbf{X}_i}+e$$

where $\mathbf{P}_{\mathbf{X}_i}$ is a projection matrix and e is the orthogonal component, which is mean zero given $\widetilde{\mathbf{X}}$.[52] This means that η_{vi} is in the space spanned by the columns of $\widetilde{\mathbf{X}}$ and hence term (C) is zero:

$$\text{(C) :} \qquad \mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\text{log density}\right)^{-1}\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\mathbb{E}\left[\eta_{vi}\middle|\,\mathbf{Z},\widetilde{\mathbf{X}},\text{log density}\right]\right]=0$$

For term (D), we assumed that ξ_{vi} was unpredictable given all unobservables in the model, so that by construction, we have:

$$\text{(D) :} \qquad \mathbb{E}\left[\xi_{vi}\,\middle|\,\mathbf{Z},\widetilde{\mathbf{X}},\text{log density}\right]=0.$$

So, we are left with the following expression for the bias in $\widehat{\theta}_{IV}$:

$$\text{Bias}\left(\widehat{\theta}_{\text{IV}}\right)=\mathbb{E}\left[\left(\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\text{log density}\right)^{-1}\mathbf{Z}'\mathbf{M}_{\widetilde{\mathbf{X}}}\;\mathbb{E}\left[c_v^U\,\middle|\,\mathbf{Z},\widetilde{\mathbf{X}},\text{log density}\right]\right]$$

So, in order for our IV estimator to be unbiased, a sufficient condition is that conditional on $\mathbf{Z}$ and the other observed individual, and village-level variables, c_v^U is mean zero.

$$\mathbb{E}\left[c_v^U\,\middle|\,\mathbf{Z},\widetilde{\mathbf{X}},\text{log density}\right]=0$$

What this amounts to is that our density shifters, namely soil quality of the village or lagged population measures, need to be uncorrelated with unobservables that influence overall village-level social capital. We discuss the plausibility of this assumption in the main text of the paper.

C.5 Mean Effects Analysis

Our estimate of the mean effects of density on groups of related outcomes is based on the procedure in Kling et al. (2007) (see footnote 22). Let $k=1,...,K$ index outcome variables for a group of related outcomes (e.g. trust in neighbors, inter-ethnic tolerance, etc.). Let $\mathbf{Y}=[\mathbf{y}_1',\mathbf{y}_2',...,\mathbf{y}_K']'$ denote a stacked vector of K social capital outcomes for that group. Let $\mathbf{X}$ be denoted as follows:

$$\mathbf{X}=\begin{bmatrix}\mathbf{X}_1 & 0 & ... & 0\\ 0 & \mathbf{X}_2 & ... & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & ... & \mathbf{X}_K\end{bmatrix}$$

[52] This is equation (11) in Altonji and Mansfield (2018).

where $\mathbf{X}_k$ consists of the control variables for outcome k. In our initial regression, this will only include individual-specific controls $\mathbf{X_i}$, predetermined community characteristics $\mathbf{C_{2v}}$ and the city-specific intercepts, but later regressions will add $\mathbf{X_v}$. Also, stack the independent variables as follows:

$$\mathbf{D} = \begin{bmatrix} \text{log density} & 0 & ... & 0 \\ 0 & \text{log density} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \text{log density} \end{bmatrix}$$

After stacking and appropriately arranging the variables, we can estimate mean effects using the following single regression:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}\theta + \varepsilon \tag{15}$$

With OLS, this is straightforward, but with IVs, we need to use an appropriately stacked vector:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & ... & 0 \\ 0 & \mathbf{Z}_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \mathbf{Z}_K \end{bmatrix}$$

After estimating the parameters of this regression (with OLS or IV), we obtain the mean effect estimate as follows:

$$\tau = \frac{1}{K}\sum_{k=1}^{K}\frac{\theta_k}{\sigma_k}$$

where σ_k is the standard deviation of $\mathbf{y}_k$. We ignore sampling variation in σ_k when estimating τ.

C.6 IFLS Panel Specification (SUR System)

Another approach to dealing with sorting is to use a two-step estimator (Combes et al., 2010). In the first step, we estimate local fixed effects of social capital, which condition out the impact of individual-specific effects and the effect of time-varying individual-level observables. We then average the residuals from this regression, and estimate a cross-sectional regression of the average social capital measures (averaged over village years) on our density measure in 2010.

Let $k = 1, ..., K$ denote a group of related social capital outcomes, and let Y_{it}^k: social capital outcome k for individual i at time t. The vector $\mathbf{x}_{it}$ consists of time-varying individual-level characteristics for individual i, such as that individual's educational attainment at time t, their marital status at time t, age at time t, and household size at time t. We first stack the outcome variables in a vector, $\mathbf{Y} = [\mathbf{y}_1', \mathbf{y}_2', ..., \mathbf{y}_K']'$. Let $\mathbf{X}$ be denoted as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 & ... & 0 \\ 0 & \mathbf{X}_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \mathbf{X}_K \end{bmatrix}$$

where $\mathbf{X}_k$ consists of the control variables for outcome k (e.g. the time-varying individual-level controls). Let $\mathbf{D}$

be a matrix individual fixed effects for outcome k (i.e. d_i), specific to each individual and outcome:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & 0 & ... & 0 \\ 0 & \mathbf{D}_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \mathbf{D}_K \end{bmatrix}$$

Let $\mathbf{A} = [\alpha_{jt}]$ denote a row vector of effects for each village and year (e.g. the α_{jt}'s). This is common to all outcomes.

In the first step of the Combes et al. (2010) estimator, we estimate α_{jt} for all equations using a SUR system:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}\theta + \mathbf{A} + \varepsilon \tag{16}$$

The $\mathbf{A}$ vector will contain the α_{jt} estimates, which are estimates of the social capital index for each village and year, after conditioning out individual fixed effects and time-varying individual-level observables.[53]

Next, in the second step, we form the cross-sectional average of these village-year fixed effects:

$$\alpha_j = \frac{1}{T}\sum_{t=1}^{T}\alpha_{jt}$$

We then use α_j as the independent variable in the following regression:

$$\alpha_j = \mathbf{C}_{2v}\beta_2 + \theta \log \text{density}_v + \Delta\varepsilon_i$$

where we instrument log density$_v$ with our 2 sets of instruments.

C.7 Heterogeneous Effects (SUR System)

To estimate heterogeneous effects of density, we add a matrix of levels terms,

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & 0 & ... & 0 \\ 0 & \mathbf{M}_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \mathbf{M}_K \end{bmatrix}$$

where $\mathbf{M}_K$ consists of the level variables for outcome k, and a stacked vector of interaction terms with density,

$$\mathbf{N} \equiv \mathbf{M}\cdot\mathbf{D} = \begin{bmatrix} \mathbf{M}_1\cdot\mathbf{D}_1 & 0 & ... & 0 \\ 0 & \mathbf{M}_2\cdot\mathbf{D}_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \mathbf{M}_K\cdot\mathbf{D}_K \end{bmatrix}$$

to (15), and we estimate the following regression using our instrument set which is augmented by interactions between $\mathbf{M}$ and the original instruments:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{M}\gamma + \mathbf{D}\theta + \mathbf{N}\theta_1 + \varepsilon \tag{17}$$

[53] Note that we can also use a single-index approach in the first step, where we form an average of the dependent variables, $Y_{it} = \frac{1}{K}\sum_{k=1}^{K} Y_{it}^k$, and just use this in a single regression to estimate α_{jt}. The results of this approach are shown in Appendix Table A.20.

Then we obtain the mean effect estimates for the reference group, τ, and for the interaction terms, τ_1, as follows:

$$\tau = \frac{1}{K}\sum_{k=1}^{K}\frac{\theta_k}{\sigma_k} \qquad \tau_1 = \frac{1}{K}\sum_{k=1}^{K}\frac{\theta_{1k}}{\sigma_k}$$

where σ_k is the standard deviation of $\mathbf{y}_k$.

We report estimates of the mean effects and the interaction terms in Table 7. Appendix Table A.21 describes how we construct different variables used in this heterogeneity analysis.