WPS7473

Policy Research Working Paper 7473

Households or Locations?

Cities, Catchment Areas and Prosperity in India

Yue Li
Martín Rama

WORLD BANK GROUP
East Asia and the Pacific Region
Office of the Chief Economist
November 2015

Policy Research Working Paper 7473

Abstract

Policy makers in developing countries, including India, are increasingly sensitive to the links between spatial transformation and economic development. However, the empirical knowledge available on those links is most often insufficient to guide policy decisions. There is no shortage of case studies on urban agglomerations of different sorts, or of benchmarking exercises for states and districts, but more systematic evidence is scarce. To help address this gap, this paper combines insights from poverty analysis and urban economics, and develops a methodology to assess spatial performance with a high degree of granularity. This methodology is applied to India, where individual household survey records are mapped to "places" (both rural and urban) below the district level. The analysis disentangles the contributions household characteristics and locations make to labor earnings, proxied by nominal household expenditure per capita. The paper shows that one-third of the variation in predicted labor earnings is explained by the locations where households reside and by the interaction between these locations and household characteristics such as education. In parallel, this methodology provides a workable metric to describe spatial productivity patterns across India. The paper shows that there is a gradation of spatial performance across places, rather than a clear rural-urban divide. It also finds that distance matters: places with higher productivity are close to each other, but some spread their prosperity over much broader areas than others. Using the spatial distribution of this metric across India, the paper further classifies places at below-district level into four tiers: top locations, their catchment areas, average locations, and bottom locations. The analysis finds that some small cities are among the top locations, while some large cities are not. It also finds that top locations and their catchment areas include many high-performing rural places, and are not necessarily more unequal than average locations. Preliminary analysis reveals that these top locations and their catchment areas display characteristics that are generally believed to drive agglomeration economies and contribute to faster productivity growth.

This paper is a product of the Office of the Chief Economist, East Asia and the Pacific Region. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at yli7@worldbank.org.

Households or Locations?
Cities, Catchment Areas and Prosperity in India

Yue Li and Martín Rama [*]

Keywords: poverty, labor earnings, location effects, spatial analysis, urbanization, catchment areas.

JEL classification: O18; I32; J31; R12; R23; C21

[*] Yue Li and Martín Rama are with the office of the Chief Economist for South Asia, at the World Bank. The authors gratefully acknowledge the skillful research assistance provided by Virgilio Galdo and María Florencia Pinto, and the useful comments and suggestions received from Urmila Chatterjee, Rinku Murgai, Ambar Narayan, and Mark Roberts. The research was partly funded by the Department for International Development of U.K. as part of the Sustainable Urban Development Multi-Donor Trust Fund.

1. Introduction

As production diversifies away from agricultural activities into manufacturing and services, the economic landscape evolves too. Urbanization is the most obvious manifestation of this change. But the spatial transformation goes beyond the emergence and growth of cities, as rural areas also densify and the boundaries between urban areas and the countryside become blurred. Policy makers in developing countries are increasingly interested in the implications of this spatial transformation. However, there is limited empirical evidence available to rigorously answer their queries. Case studies about specific cities abound, and there is also a wealth of benchmarking exercises across different administrative levels, including metropolitan areas, states or districts. There are also lessons from urban studies conducted in advanced economies, where urbanization was completed decades ago. But there are few systematic studies on the contribution the rural-urban transformation makes to economic growth and poverty reduction in countries that are still in the process of urbanizing.

Much of the available evidence on the relationship between locations and prosperity in developing countries comes from poverty analysis, and especially from the literature on poverty maps. These poverty maps provide a succinct measure of average household expenditure (or income) per capita in real terms across space, at a fairly disaggregated level. Building on theories of consumption they use household surveys, whose samples are small but rich in information, to estimate the relationship between household expenditure per capita and household characteristics. The set of characteristics considered are those that can also be found in population censuses. The estimated relationship is then used to predict household per capita expenditures at disaggregated spatial levels, based on local household characteristics as reported by population censuses (Demombynes et al. 2002, Elbers, Lanjouw and Lanjouw 2003, Hentschel, Lanjouw et al. 2000).

Despite the use of the word “map”, these analyses remain focused on using household characteristics to predict household expenditure, rather than on understanding location effects. Some location characteristics are generally introduced in the empirical analysis, but this is mainly to reduce biases in the prediction of household expenditures per capita. Efforts to unpack the contribution locations make to poverty prevalence have remained fairly aggregated, using the region or the province as the spatial unit of analysis, or distinguishing between urban and rural areas taken altogether (Kanbur and Venables, 2005).

Admittedly, this strand of literature includes analyses of the growth in household expenditures per capita which explicitly focus on local “poverty traps”. The use of panel data in these analyses allows controlling for unobservable household characteristics which could be spatially correlated, and whose impact could therefore be wrongly construed as a location effect. The analyses also introduce a range of location characteristics at the fairly disaggregated levels, including topography, remoteness, density of rural roads, and local human development indicators. Many of these characteristics are shown to contribute significantly to the growth in household expenditures per capita, which is interpreted as evidence that geographic capital can influence the productivity of a household's own capital (Ravallion and Jalan 1999, and Jalan and Ravallion 2002). But these analyses are restricted to farm households in rural areas, so that they are more informative about bottom locations than about the broader rural-urban transformation.

Urban economics, on the other hand, squarely focuses on cities. This other strand of literature aims to quantify agglomeration economies, as reflected in spatial disparities in nominal wages. Underpinned by theories of local externalities, its basic premise is that firms and workers are more productive in large and dense urban environments (Rosen 1979, and Roback 1982). The analyses emphasize location characteristics perceived as being directly relevant to the strength of such local externalities, including population size, population density, and employment density (Combes, Duranton, and Gobillon 2008, Combes et al. 2010, Glaeser and Maré 2001). Location characteristics are also highlighted in connection to the potential channels underpinning agglomeration economies. For example, locations may be more productive because of knowledge spillovers, in which case the level of local skills is a variable of interest (Rauch 1993, Moretti 2004a and 2004b, Rosenthal and Strange 2008). Other location characteristics usually considered are natural resource endowments and climate (see Duranton 2014, Gill and Goh 2009, Glaeser and Gottlieb 2009, Puga 2010, and Rosenthal and Strange 2004 for reviews).

In the context of advanced economies, urban economics has made important progress in identifying the implications of location characteristics for employment and pay, and for economic development more generally. However, the urban economics approach on its own may also be insufficient to fully understand the implications of the rural-urban transformation in developing countries. Its unit of observation is typically the city, which leaves out not only the rural areas where a large fraction of the population still lives, but also the increasingly blurred areas at the urban fringe. There are also important data limitations, as only a minority of workers in developing countries are wage earners, and data on their nominal earnings are often partial and unreliable (World Bank, 2012).

These two strands of literature have so far developed largely disconnected from each other, to the point that studies belonging to one of them are rarely cited in studies from the other. Both poverty analysis and urban economics use disaggregated household or individual data to predict an indicator of expenditure or income, but they do so very differently. And yet, in their different ways these two analytical bodies are dealing with same issue, namely taking into account the role of location in explaining prosperity.

In this paper, we draw insights from the two strands of literature and develop a hybrid methodology to assess spatial performance with a high level of granularity. As in urban economics, we are interested in the spatial distribution of labor productivity. Earnings from labor are indeed the most important component of household income in developing countries. But given the limitations of wage data when a majority of workers are farmers or self-employed, we approximate labor earnings through nominal household expenditures per capita, as in poverty analysis. A key element of our methodology is to conduct the analysis across all locations, regardless of whether they are administratively urban or rural.

We illustrate this methodology in the case of India. This is the country with the largest number of poor people, worldwide (World Bank 2015). It is also a country at an early stage in the urbanization process, where regular wage workers only account for 18 percent of the labor force and information on wages or labor earnings is available for only 45 percent of it (NSSO 2012). These characteristics make India ideally suited to combine insights from urban economics and poverty analysis. Moreover, the nature of the available household survey data allows us to generate estimates with fairly high spatial disaggregation. Building on an approach developed by Chatterjee et al. (2015) we can indeed distinguish between small rural, large rural, small urban and large urban areas within each district. While not all developing countries have household survey data supporting such level of granularity, we believe that the methodology proposed in this paper can be applied to other country settings and yield insights about their own rural-urban transformations.

Our results confirm that location is an important determinant of labor productivity, even after controlling for a wide range of household characteristics. In India's case, one third of the variation in predicted labor earnings is explained by the locations where households reside and the interactions between locations and household characteristics. Importantly, this methodology provides a reasonable metric to systematically describe spatial productivity patterns across the entire country. On average, large rural areas perform better than small rural areas, and large urban areas perform better than small urban areas. But the performance of large rural areas and that of small urban areas resemble closely, challenging the conventional view of a rural-urban divide. We also find that performance is spatially correlated. Places with higher productivity tend to locate close to each other, and so do places with lower productivity. The spatial correlation attenuates over distance. However, "distance to what?" is important as well. Some high-performance places spread their prosperity over much broader areas than others.

The importance of distance, and especially of "distance to what?" suggests that places should not be looked at independently from each other. We use this insight to further classify all places into four tiers: top locations, their catchment areas, average locations and bottom locations. The classification relies on the distribution of the performance metric generated by our methodology and on the distance between places. It results in the identification of 17 clusters of top locations and their catchment areas across India. These clusters include many high performing rural places, and their better performance is not necessarily associated with higher levels of inequality. Based on the classification, we also report the correlations between the factors that potentially drive agglomeration economies, or contribute to faster productivity growth, and the tier that a location falls into.

2. Poverty analysis meets urban economics

Both poverty analysis and urban economics try to explain the variation in expenditure or income within a country, and both consider a spatial dimension in that variation. They typically do so by introducing location effects in their empirical work. But spotting the nuances in the way they do it is important to find a common methodological ground between them.

In poverty analysis, the variable of interest is real household expenditure per capita and the key explanatory variables are household characteristics such as size, composition by age and gender, educational attainment, asset ownership and the like. Location characteristics such as topography, distance to markets, or the availability of basic services, are often included in the analysis. When constructing poverty maps, cluster-specific disturbances are introduced to account for the potential correlation between unobservable household characteristics living in the same geographic area, which could bias the estimates. Thus, the typical empirical specification takes the form:

$$Ln\left(\begin{matrix}Real\ expenditure\\ per\ capita\end{matrix}\right)_{hl} = \alpha^P + \beta^P.\left(\begin{matrix}Household\\ characteristics\end{matrix}\right)_{hl} + \theta^P.\left(\begin{matrix}Location\\ characteristics\end{matrix}\right)_{l} + \mu_l^P + \varepsilon_{hl}$$

where h denotes households, l denotes locations, μ_l^P is the cluster-specific disturbance, and ε_{hl} is an error term whose mean is equal to zero.

While poverty assessments typically build on some variation of this equation, the construction of poverty maps follows a more structured empirical strategy to select the most relevant location characteristics. First, the equation above is estimated without including any location characteristics in the specification. Then, the resulting cluster-specific fixed effects are regressed on a broad range of location characteristics. In the final step, the location characteristics displaying the best fit are introduced, together with household characteristics, in the regression.

Urban economics, on the other hand, uses the equivalent of an augmented Mincerian equation to quantify agglomeration economies. The variable of interest is the nominal wage. Based on human capital theory, the key explanatory variables are the workers' educational attainment and work experience, generally proxied by age. What urban economics adds is a set of location characteristics which are supposedly associated with stronger agglomeration effects. Examples of such location characteristics include population size, population density, connectivity, sectoral structure of production, and average skills. The typical specification in this case is:

$$Ln\binom{Nominal}{wage}_{ic} = \alpha^U + \beta^U.\binom{Worker}{characteristics}_{ic} + \theta^U.\binom{Location}{characteristics}_{c} + \varepsilon_{ic}$$

where *i* denotes individuals and *c* denotes cities (a subset of all locations *l*).

As in the case of poverty maps, a multi-step empirical strategy has been adopted by some studies. In the first step, a regression of individual nominal wages on worker characteristics and city fixed-effects is estimated. In the second step, the estimated city fixed effects are regressed on city characteristics that influence agglomeration economies or capture the channels underpinning those effects, as well as on other factors that may affect local wages. This two-step approach allows to disentangle the contribution worker characteristics and location characteristics make to the spatial distribution of wages.

Finding a common ground between these two approaches requires clarity on the relationship between their respective variables of interest. Nominal wages are a reasonably reliable indicator of labor productivity. From the workers' point of view, higher nominal wages may not necessarily reflect higher living standards, as large and dense urban environments are also characterized by higher rents, more expensive goods and services, and congestion costs. But firms would only be willing to pay these higher nominal wages if workers in these locations were more productive. Everything else equal, firms producing goods that are traded nationally would select to locate in high-wage places only if the local productive advantage is significant. As long as there are some firms producing traded goods in every place, average productivity needs to be higher in places where nominal wages are higher (Acemoglu and Angrist 2000). And as long as labor markets are relatively efficient, higher nominal wages should be correlated with higher labor earnings among workers who are not wage earners.

Variation in labor earnings in turn drives variation in household expenditures, but the two variables are not perfectly correlated. On the income side, some households also generate income from assets such as land, and some receive remittances or social assistance transfers. On the consumption side, the same labor earnings can result in very different levels of expenditure per capita depending on the household's size. The relationship between labor earnings and expenditure per capita is also shaped by preferences and norms, as they influence savings rates.

Some of the gaps between the two variables of interest can be attributed to the household themselves, while others are to a larger extent due to the location where the households live. Controlling for household characteristics such as size and age composition allows to account for different denominators when reporting labor earnings on a per capita basis. Controlling for household assets arguably takes care of non-labor incomes. And controlling for social background and religion goes some way towards introducing household preferences and norms. On the other hand having migrant members, commuting for work, or receiving social assistance transfers is arguably influenced by location characteristics, such as the reach of social protection systems and the availability of job opportunities at the local level.

Our benchmark specification is inspired by the first step of the empirical strategy considered by both poverty maps and urban economics:

$$Ln\left(\begin{matrix}Nominal\ expenditure\\ per\ capita\end{matrix}\right)_{hl} = \alpha + \beta.\left(\begin{matrix}Household\\ characteristics\end{matrix}\right)_{hl} + \gamma_l + \varepsilon_{hl}$$

where γ_l are location effects. Ideally, this equation should also include household effects to control for unobservable household characteristics, such as work ethic or entrepreneurial spirit. But doing so would require panel data, which is not available in India's case. While being aware that these unobservable household characteristics could bias the estimates, we believe that the risk is mitigated by the use of a large number of observable characteristics among the explanatory variables of the regression. If this is correct, the estimated location effects should provide a reasonably good approximation to the spatial variation in productivity, hence to the magnitude of agglomeration economies across the country.

Our approach allows us to disentangle the contribution household characteristics and location effects make to labor earnings. This understanding is highly relevant from a policy perspective. Implicit in the approach is the assumption that households make the most of both their assets and the opportunities provided by the places where they live. Educational investments, occupational choices and migration decisions (either permanent or seasonal) are shaped by the interaction between household characteristics and location characteristics. But this interaction is somewhat overlooked by traditional poverty analysis as it emphasizes households over locations, and focuses its recommendations on upgrading skills and other household assets, or on better targeting resource transfers to the poor. These interventions are certainly important, but there may be a need to rebalance development priorities and bring more attention to local externalities—both positive and negative – affecting household choices.

Further, the approach allows us to get a reasonable assessment of the spatial variations of productivity without being too constraining on the underlying mechanisms. Many channels have been highlighted as potential sources of agglomeration economies, including the pooling of labor, the sharing of resources and productive amenities, reductions in transportation costs, and knowledge spillovers (Marshall 1890, Jacobs 1969, Krugman 1991). We see exploring these channels carefully together with other local factors as the next step in our research agenda. But as a first step, location effects provide us a workable metric to describe spatial productivity patterns across India.

A key element of our methodology is to conduct the analysis across all locations, regardless of whether they are administratively urban or rural. Many poverty analyses focus on rural areas, because that is where a majority of the poor live. Urban economics, on the other hand, focuses on cities and towns. But in a

rapidly urbanizing country, like India, the boundaries between rural and urban areas are often blurred. According to the Census of India 2011, 3,894 villages are administrative rural but display economic characteristics closely resembling those of cities (Office of the Registrar General and Census Commissioner 2011a). A special name, census town, has even been coined to label this gray area in the rural-urban gradation. In fact, about 30 percent of India's urbanization between 2001 and 2011 is attributable to the reclassification of rural areas as census towns (Pradhan 2013). By considering all locations, our approach avoids the pitfalls from somewhat arbitrary administrative classifications.

3. The empirical strategy

Implementing the approach outlined above requires information on individual households as well as a robust mechanism to match each household observation to a particular location. Characterizing the locations, say in terms of their connectivity, also requires spatial data.

The household survey data used in this paper is from the Schedule 1.0, Household Consumer Expenditure Survey of the 68th round of National Sample Survey of India conducted between July 2011 and June 2012 (NSSO, 2012). This survey, hereafter identified as NSS 2011-12, reports household consumption information on an itemized form, based on a 30-day recall period, and based on a mixed recall period. We use monthly consumption based on the mixed recall period and household size to compute monthly per capita nominal household expenditure, which is the explained variable in our benchmark specification.

The survey also reports demographic data, educational attainment, landholdings, the source of energy used for lighting and for cooking, the social group the household belongs to, and its religious affiliation. We use this information to construct the household characteristics of the benchmark specification. It must be noted that the monthly consumption reported by the NSS includes paid rent but excludes imputed rent. This biases downward the expenditure of households who own their dwelling units. To address this data limitation, we add information of household dwelling ownership to the set of household characteristics in the analysis.

As for locations, most poverty analyses in India consider the state or the region as the spatial unit of analysis, further dividing each unit into urban and rural areas. However, that level of aggregation is too high to assess local externalities. Even data at the district level may not be disaggregated enough for that purpose. Separately identifying individual cities would be difficult too, because the actual boundaries of urban agglomerations do not match well the administrative boundaries and classifications used by household surveys (Li et al., 2015). To address these limitations, in this paper we adopt the approach developed by Chatterjee et al. (2015) to generate estimates below district level. Their approach uses the design of the NSS 2011-12 to estimate the population of first-stage sampling units. The resulting characterization of the employment structure across these different population sizes challenges the conventional wisdom of a rural-urban "divide" in India's context.

The NSS 2011-12 covers all of India except interior villages of Nagaland situated beyond five kilometers of a bus route and villages in Andaman and Nicobar Islands. The survey follows a stratified multi-stage sampling design. Each district of a state or union territory is stratified into rural and urban areas. In the rural stratum, the first stage units are the 2001 census villages; in the urban stratum, they are urban frame survey blocks. Within each stratum, first-stage units are ordered by their population and then further stratified. The ultimate stage units are households, drawn from the selected first stage units of each substrata.

Following Chatterjee et al. (2015), we classify the first-stage units (villages or urban frame survey blocks) of each district into four groups, based on the average population size of their substratum. The four groups are: 1) small rural areas with a population less than 5,000; 2) large rural areas with a population above 5,000; 3) small urban areas with a population less than one million; and 4) large urban areas with a population greater than one million. Breakdowns of this sort are not unusual in urban economics (e.g. Glaeser and Maré 2001). In what follows we use the word "place" to refer to a population size group within a district, and interpret the location subscript *l* in our benchmark specification as referring to places. In most cases, a place includes more than one first-stage unit. In the case of large urban areas and some small urban areas, a place can be interpreted as corresponding to a city.

In principle, each district could include first-stage units belonging to all four population size groups. But in reality not all districts host large urban areas, or even small urban areas; some do not even include first-stage units in the large rural category. Also, because of limited information, population size ranges cannot be estimated for the union territory of Andaman and Nicobar Islands, the union territory of Daman and Diu, and the state of Nagaland (see Chatterjee et al. 2015 for details). Our analysis excludes the island state of Lakshadweep, due to problems with the measurement of distance between districts, a variable needed in the analysis below. Furthermore, to reduce measurement errors caused by the mismatch between the districts defined by NSS 2011-12 and by the Census 2011, we merge all observations from the union territory of Delhi into one district.

As a result of these constraints and adjustments, our analysis covers 1,406 places from 599 districts in 31 states or union territories. Among them, 579 are small rural areas, 221 are large rural areas, 581 are small urban areas, and 25 are large urban areas. Our final sample includes the 96,227 households in the sample who live in the 1,406 places retained and who report information on all the household characteristics considered in the analysis.

In order to incorporate the geographic distribution of these places in our analysis, we digitized the administrative boundaries into a standard digital vector storage format for spatial data, or shapefile, relying on the Administrative Atlas of India 2011 (Office of the Registrar General and Census Commissioner 2011b). Based on NSS 2011-12, first-stage units can be identified down to the district level. Unfortunately, there is not sufficient information for us to go further down, to the tehsil level. Therefore, the digitization is conducted at the district level.

There are also inconsistencies in the spatial framework of the NSS and the Atlas which require further adjustments. The Atlas contains 640 district-level polygons corresponding to the 640 districts defined by the Census of India 2011 (Office of the Registrar General and Census Commissioner 2011a). However, the NSS 2011-12 only includes 621 districts, because its sample frame is based on the administrative boundaries defined by the Census of India 2001 (Office of the Registrar General and Census Commissioner 2001). We match the 621 districts in the NSS 2011-12 to the 640 districts of the Census of India 2011 using information available from the Atlas, the official websites of the districts, and other relevant sources. We restrict our analysis to the districts that exist in both NSS 2011-12 and the Census of India 2011. Because we combine observations of Delhi into one district, we further merge the district level polygons of Delhi into one. Similarly, because NSS 2011-12 subsumes the district of Mumbai into the district of Mumbai suburban, we merge the polygons of these two districts as well. Finally, we do not consider districts that emerged after the NSS 2011-12 defined its sample frame.

Given that digitization of administrative boundaries is at the district level, computing the distance between places requires additional assumptions. For places belonging to different population size groups but within the same district, the distance is assumed to be zero. For places in different districts, regardless of their population size group, we use the pairwise distance between the corresponding districts. For any two districts, the pairwise distance is computed as the length of the shortest surface-level curve between their centroids, based on the Haversine formula.

We also generate information on the location characteristics of individual places, using to that effect the Spatial Database for South Asia (Li et al., 2015). This platform combines data from the Census of India, the Household Consumer Expenditure and the Employment and Unemployment modules of the NSS, the Economic Census, administrative records, remote sensing data and crowdsourced data. The Spatial Database for South Asia provides information on a range of socioeconomic indicators, including the urban extent, demographics, jobs, economic activity, infrastructure, ICT, finance, business, living standards, education, health, and environment (Table 1).

Table 1. Summary statistics, by type of location

	Small rural		**Large rural**		**Small urban**		**Large urban**		**Total**	
Places	579		221		581		25		1406	
Observations at household level	45873		10785		33651		5918		96227	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Household expenditure per capita (current India Rupees per month)	1527	1133	2063	2736	2466	2246	3590	3277	2042	2052
Demographics										
Household size	4.94	2.23	4.70	2.16	4.41	2.16	4.15	2.20	4.68	2.21
Children under 6	0.07	0.10	0.07	0.10	0.06	0.10	0.05	0.10	0.06	0.10
Children above 6	0.10	0.11	0.10	0.11	0.09	0.11	0.08	0.11	0.10	0.11
Female adults	0.23	0.13	0.25	0.14	0.24	0.15	0.23	0.16	0.24	0.14
Female dependents	0.03	0.08	0.04	0.09	0.03	0.08	0.03	0.08	0.03	0.08
Male dependents	0.04	0.09	0.04	0.10	0.03	0.09	0.03	0.09	0.04	0.09
Female household head	0.09	0.29	0.13	0.34	0.12	0.33	0.10	0.30	0.11	0.31
Skills										
Maximum education of adults (years)	8.40	4.46	9.02	4.41	10.45	4.46	11.06	4.44	9.35	4.57
Assets										
Land (0.000 hectares)	0.96	2.14	0.53	1.41	0.19	1.03	0.06	0.64	0.59	1.72
Dwelling										
Own	0.95	0.21	0.93	0.26	0.69	0.46	0.58	0.49	0.84	0.37
Rent	0.03	0.17	0.06	0.23	0.27	0.44	0.37	0.48	0.14	0.34
Other	0.02	0.13	0.02	0.13	0.04	0.19	0.05	0.21	0.03	0.16
No	0.00	0.03	0.00	0.02	0.00	0.03	0.00	0.05	0.00	0.03

(Continued)

Table 1. Summary statistics, by type of location (continued)

	Small rural		Large rural		Small urban		Large urban		Total	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Housing										
Energy for cooking										
Coke and coal	0.01	0.11	0.01	0.08	0.03	0.16	0.01	0.08	0.02	0.13
Firewood and chips	0.66	0.47	0.53	0.50	0.20	0.40	0.04	0.19	0.45	0.50
LPG	0.22	0.41	0.35	0.48	0.68	0.47	0.78	0.41	0.43	0.49
Gobar gas	0.00	0.05	0.00	0.05	0.00	0.02	0.00	0.00	0.00	0.04
Dung cake	0.07	0.26	0.07	0.25	0.01	0.12	0.00	0.07	0.05	0.21
Charcoal	0.00	0.02	0.00	0.02	0.00	0.05	0.00	0.01	0.00	0.03
Kerosene	0.01	0.09	0.01	0.10	0.04	0.18	0.08	0.27	0.02	0.15
Electricity	0.00	0.03	0.00	0.03	0.01	0.08	0.00	0.05	0.00	0.06
Other	0.03	0.16	0.04	0.18	0.01	0.07	0.03	0.16	0.02	0.14
No	0.00	0.05	0.01	0.08	0.03	0.16	0.06	0.24	0.01	0.12
Energy for lighting										
Kerosene	0.19	0.39	0.15	0.35	0.04	0.20	0.01	0.10	0.12	0.33
Other oil	0.00	0.02	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.02
Gas	0.00	0.03	0.00	0.04	0.00	0.03	0.00	0.04	0.00	0.03
Candle	0.00	0.05	0.00	0.05	0.00	0.05	0.00	0.03	0.00	0.05
Electricity	0.80	0.40	0.84	0.36	0.95	0.21	0.99	0.11	0.87	0.34
Other	0.00	0.05	0.00	0.03	0.00	0.04	0.00	0.04	0.00	0.04
No	0.00	0.07	0.00	0.05	0.00	0.04	0.00	0.02	0.00	0.05
Social and religious backgrounds										
Scheduled Tribe	0.19	0.39	0.04	0.21	0.09	0.28	0.03	0.17	0.13	0.33
Scheduled Caste	0.17	0.38	0.17	0.38	0.13	0.34	0.14	0.35	0.16	0.36
Other Backward Caste	0.38	0.49	0.49	0.50	0.41	0.49	0.30	0.46	0.40	0.49
Hindu	0.78	0.41	0.73	0.44	0.75	0.43	0.81	0.40	0.76	0.42
Muslim	0.11	0.31	0.18	0.39	0.15	0.36	0.14	0.34	0.13	0.34

4. Main results

We estimate the benchmark specification using both Ordinary Least Squares (OLS) and Weighted Least Squares (WLS). For the latter, we apply the sample weights at the household level provided by the NSS 2011-12, which ensure that the full data are representative for India. There is considerable debate on whether using OLS or WLS is preferable. No doubt, weighted summary statistics present a representative picture for the underlying population when survey data is used. But when it comes to regression analysis, WLS does not necessarily generate more consistent or more efficient estimators than OLS. Fortunately, the two methods yield very similar coefficients (Table 2).

To check whether the difference between the estimators from the two approaches is statistically significant we first apply a test described by Deaton (1997). The test consists of running a weighted regression of the OLS residuals on all the explanatory variables, and evaluating whether the estimated coefficients are jointly equal to zero. The resulting F statistic is 1.36, which is significant at the 0.01 level. However, the R-square of the regression is only 0.041, suggesting a limited difference in explanatory power between the two approaches. We further check the correlation between the parameters estimated with OLS and WLS (Figure 1). For the parameters on household characteristics the correlation coefficient is 0.98; for location effects it is 0.95. Despite these similarities, we conduct the analysis using both OLS and WLS and systematically verify that the conclusions are not dependent on the estimation method. For brevity, in what follows we only present results based on OLS estimators. Results based on WLS estimators are available upon request.

In applying OLS and WLS we implicitly assume that the error terms in the benchmark specification are independently distributed across households and locations. However, the literature on spatial econometrics shows that observations from nearby locations often exhibit similar properties and tend to be spatially correlated. This spatial correlation raises problems similar to those created by the serial autocorrelation of residuals in time-series analysis (Anselin 2003, and Anselin and Rey 2010).

To assess whether there is spatial autocorrelation in our data we run several tests on the residuals of the benchmark specification. First, we average these residuals across the 1,406 places and confirm that the mean residuals by place are distributed closely around zero. This indicates that there is no clustering of residuals at the place level. We then compute the correlation coefficients between mean residuals by place, for all size groups. We do this within each district, and also across districts at distance intervals of 50 kms, up to a maximum distance of 400 km. The resulting correlation coefficients turned out to be small and mostly insignificant (Figure 2). In the case of small rural, large rural and small urban areas, 23 of the 24 correlation coefficients between mean residuals for the same population size groups are statistically insignificant. The only exception is for small urban areas that are distant from each other between 0 and 50 km where the coefficient is around 0.19 and statistically significant. The cross-correlations of group means of residuals that belong to different population size groups shows a similar pattern: 80 of the 81 coefficients are statistically insignificant. In the case of large urban areas, the standard deviations of the correlation coefficients are much larger, because there are much fewer observations. But a vast majority of the correlation coefficients are statistically insignificant.

The lack of spatial autocorrelation of residuals implies that the results of the benchmark regression are not biased, but it does not imply a lack of spatial correlation in household expenditures per capita. Similar to the procedure used for the mean residuals by place, we compute the correlation coefficients between location effects, both within the same district and across districts, at intervals of 50 km. In sharp contrast to

what was observed for the mean residuals, the spatial correlations among location effects are strong and statistically significant (Figure 3). In the case of small rural, large rural and small urban areas, all correlation coefficients are above 0.4 and significantly different from zero for places within the same district. The correlation coefficients gradually decline for districts further apart, but they remain statistically significant for at least 200 km. In the case of large urban areas, the correlation coefficients follow a similar pattern for correlation with places belonging to other population size groups but are more volatile for the correlation coefficient with other large urban areas, because there are few of them.

Table 2 Benchmark regression results

	OLS	WLS
Location effects		
Place	Yes	Yes
Household characteristics		
Demographics		
Household size	-0.567***	-0.523***
	(0.005)	(0.009)
Children under 6	-0.384***	-0.319***
	(0.013)	(0.020)
Children above 6	-0.014	0.008
	(0.012)	(0.019)
Female adults	0.068***	0.040*
	(0.014)	(0.023)
Female dependents	-0.109***	-0.092***
	(0.018)	(0.027)
Male dependents	0.080***	0.071***
	(0.017)	(0.025)
Female household head	-0.064***	-0.055***
	(0.005)	(0.007)
Skills		
Maximum education	0.006***	-0.003*
	(0.001)	(0.001)
Maximum education squared	0.002***	0.002***
	(0.000)	(0.000)
Assets		
Land	0.179***	0.149***
	(0.003)	(0.005)
Dwelling (omitted = Own)		
Rent	0.107***	0.130***
	(0.005)	(0.008)
Other	-0.113***	-0.094***
	(0.009)	(0.018)
No	0.058	0.039
	(0.047)	(0.052)

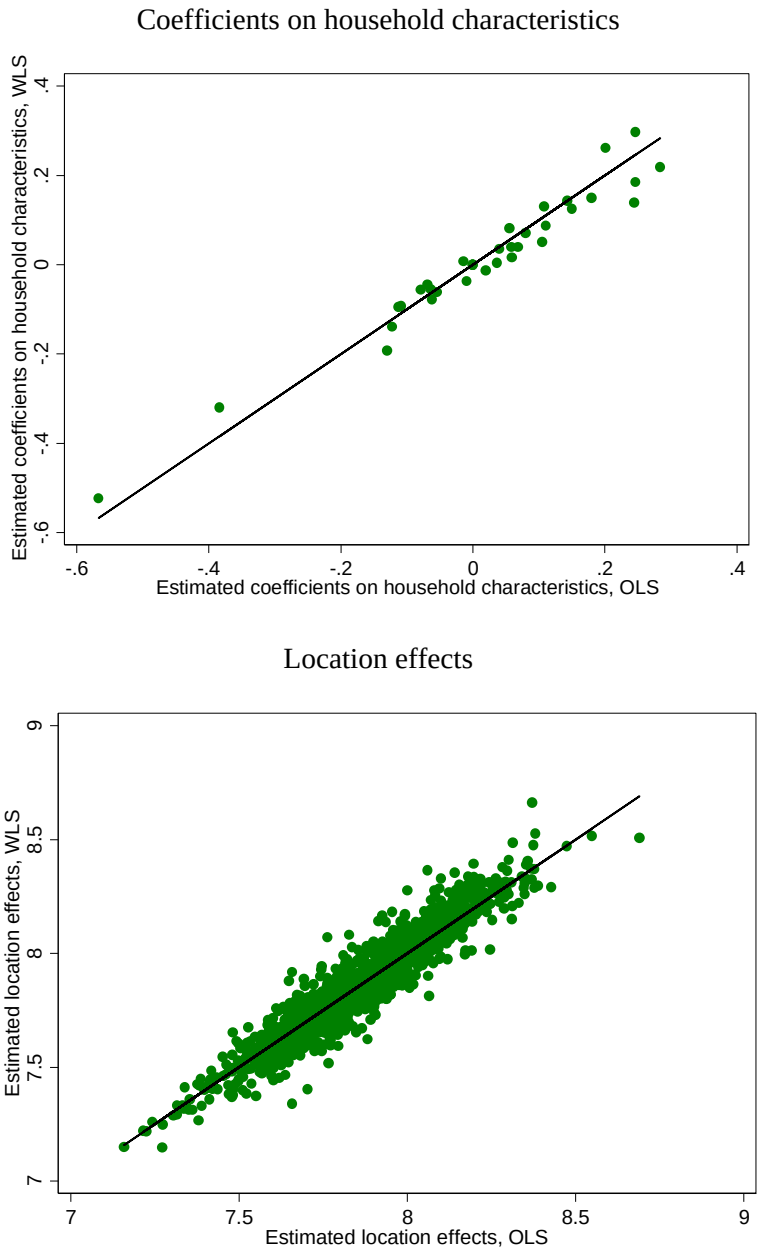
Note: Estimated coefficients *significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

(Continued)

Table 2 Benchmark regression results (Continued)

	OLS	WLS
Housing		
Energy for cooking (omitted = Coke and coal)		
Firewood and chips	0.019*	-0.013
	(0.011)	(0.016)
LPG	0.284***	0.219***
	(0.011)	(0.016)
Gobar gas	0.244***	0.139***
	(0.038)	(0.041)
Dung cake	0.059***	0.016
	(0.013)	(0.018)
Charcoal	0.143***	0.143**
	(0.034)	(0.057)
Kerosene	-0.009	-0.037*
	(0.014)	(0.022)
Electricity	0.201***	0.261***
	(0.031)	(0.063)
Others	0.037**	0.004
	(0.015)	(0.022)
No	0.246***	0.298***
	(0.019)	(0.033)
Energy for lighting (omitted = Kerosene)		
Other oil	0.111**	0.087
	(0.054)	(0.076)
Gas	0.246***	0.185***
	(0.038)	(0.062)
Candle	0.105***	0.051
	(0.027)	(0.042)
Electricity	0.150***	0.126***
	(0.004)	(0.007)
Others	-0.079*	-0.056
	(0.046)	(0.069)
No	0.055**	0.082
	(0.025)	(0.059)
Social and religious backgrounds		
Scheduled Tribes	-0.130***	-0.192***
	(0.006)	(0.010)
Scheduled Castes	-0.122***	-0.139***
	(0.004)	(0.007)
Other Backward Castes	-0.062***	-0.078***
	(0.004)	(0.006)
Hindu	-0.069***	-0.044***
	(0.007)	(0.011)
Muslim	-0.054***	-0.062***
	(0.008)	(0.012)
Observations	96227	96227
R2	0.622	0.683
R2 Adjusted	0.616	0.678

Figure 1. Correlation between OLS and WLS estimates



Note: The solid line has a 45-degree slope, corresponding to the case where OLS and WLS estimates are identical.

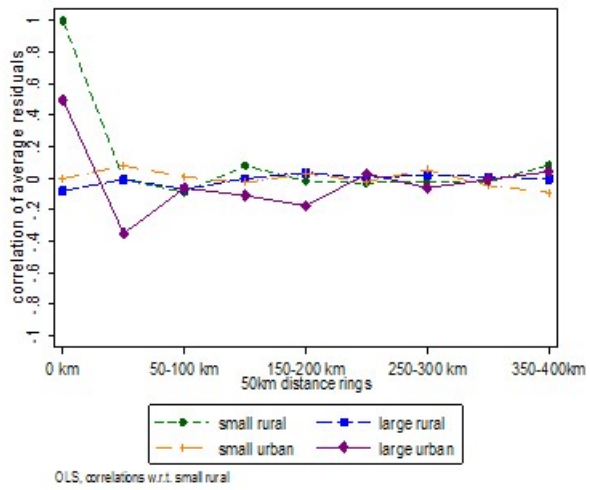
5. Location matters

Combining poverty analysis with urban economics changes the assessment of the relative contribution of household characteristics and location effects to prosperity. Adding more explanatory variables to a

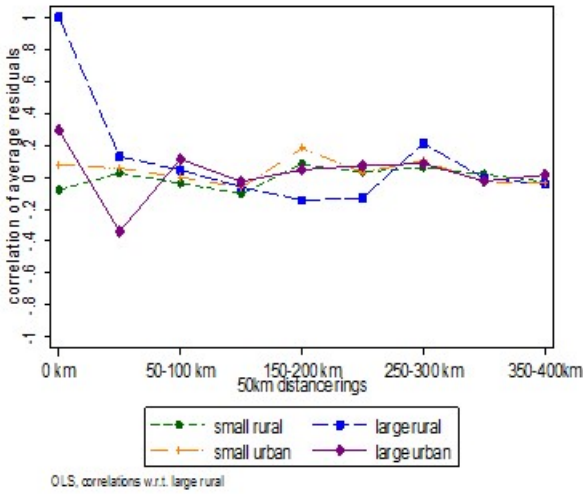
regression always increases the overall explanatory power of the model, but the increase is substantial in this case (Table 3). As a comparator to our benchmark specification, we estimate a model with only household characteristics. The household expenditure predicted by this model explains 51.5 percent of the overall variation in observed household expenditure. By contrast, the predicted expenditure of our benchmark specification with location effects at the place level explains 62.2 percent of the overall variation. For comparison purposes we also conduct two other regressions, with location effects defined at the state and the district levels.

Figure 2. Spatial correlation between average residuals by place

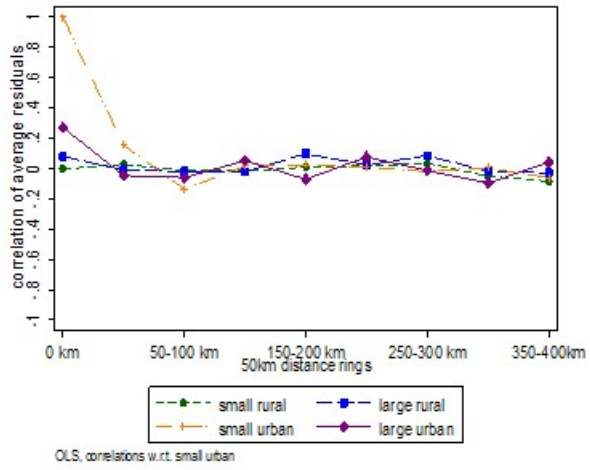
Small rural (with itself and others)



Large rural (with itself and others)



Small urban (with itself and others)



Large urban (with itself and others)

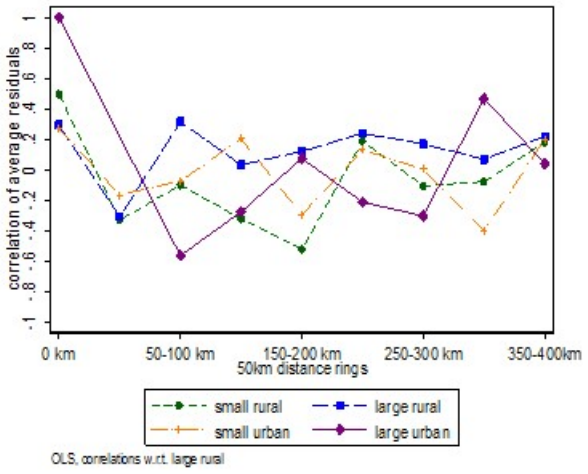
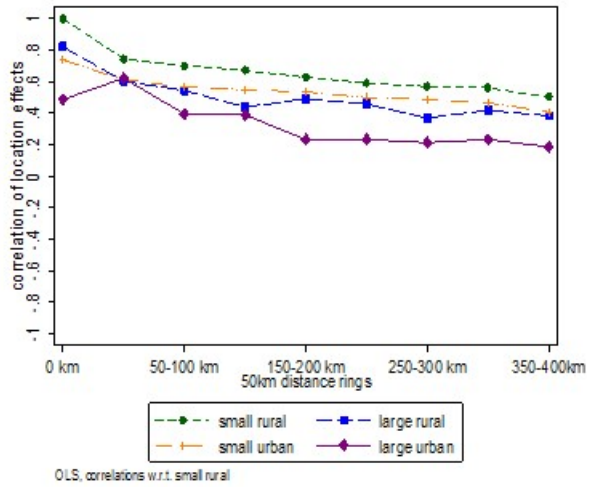
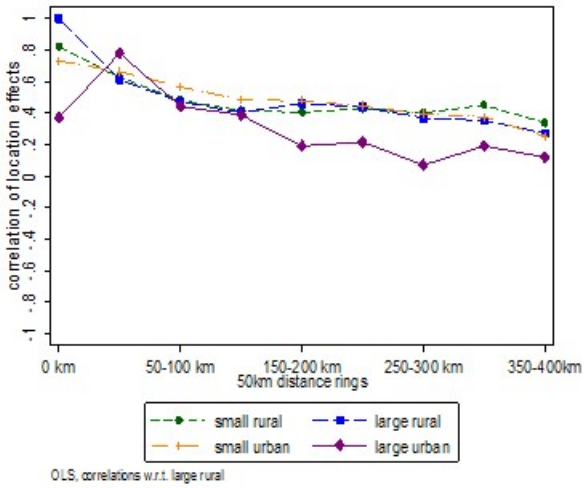


Figure 3. Spatial correlation between location effects

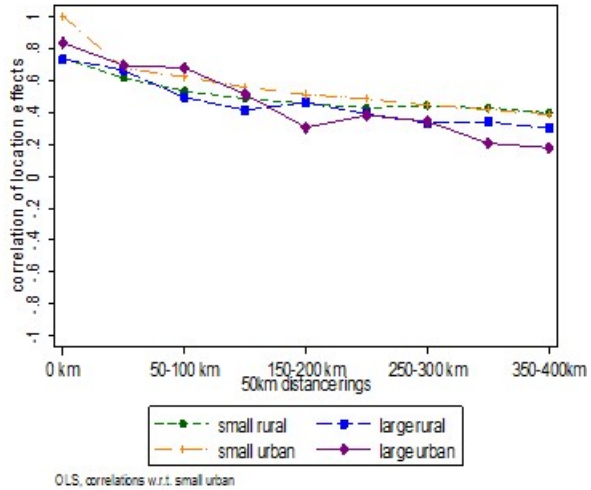
Small rural (with itself and others)



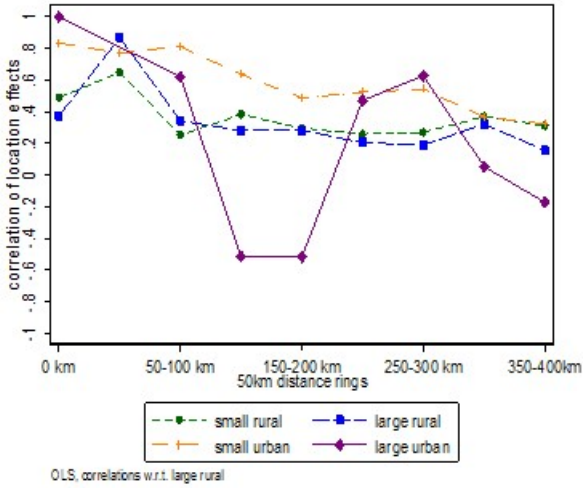
Large rural (with itself and others)



Small urban (with itself and others)



Large urban (with itself and others)



Introducing disaggregated location effects not only improves the overall fit of the model: it also corrects biases in the estimated returns to household characteristics and highlights the correlation between those characteristics and location effects. An intuitive explanation of this correlation is the sorting of households through migration decisions. Cities do not attract a random sample of the rural population, but rather specific population subsets, such as people whose educational attainment is above average. Migration is not the only mechanism at play. Cities, especially functional ones, also make people with the same characteristics, such as educational attainment, more productive (Moretti 2004a and 2004b). Conversely, socially disadvantaged groups tend to concentrate in some of the least productive places. For example, households belonging to Scheduled Tribes often live in forest areas in India. Not taking this sorting

explicitly into account results in overstating the negative impact of their social background on their household expenditure per capita.

Table 3. Variance decomposition

Model (OLS)	**1**	**2**	**3**	**4**
Location effects				
None	Yes			
State		Yes		
District			Yes	
Place				Yes
Variance				
Observed expenditure	0.388	0.388	0.388	0.388
Predicted expenditure	0.200	0.217	0.236	0.241
Household	0.200	0.176	0.165	0.153
Location	0.000	0.020	0.041	0.050
Interaction	0.000	0.021	0.030	0.038
Percentage of total variance				
Observed expenditure	100.0	100.0	100.0	100.0
Predicted expenditure	51.5	56.0	60.8	62.2
Household	51.5	45.4	42.5	39.5
Location	0.0	5.1	10.6	12.9
Interaction	0.0	5.4	7.7	9.8

Note: From our benchmark specification it follows that:

$$Var\left(ln\left(\begin{matrix}Nominal\ expenditure\\ per\ capita\end{matrix}\right)_{hl}\right) = \underbrace{Var\left(\alpha + \beta \cdot \left(\begin{matrix}Household\\ characteristics\end{matrix}\right)_{hl} + \gamma_l\right)}_{variance\ of\ predicted\ expenditure} + \underbrace{Var(\varepsilon_{hl})}_{variance\ of\ residuals}$$

$$= \underbrace{Var\left(\beta \cdot \left(\begin{matrix}Household\\ characteristics\end{matrix}\right)_{hl}\right)}_{household} + \underbrace{Var(\gamma_l)}_{location} + \underbrace{2 * Cov\left(\beta \cdot \left(\begin{matrix}Household\\ characteristics\end{matrix}\right)_{hl}, \gamma_l\right)}_{interaction} + \underbrace{Var(\varepsilon_{hl})}_{variance\ of\ residuals}$$

This point can be illustrated in a more formal way by decomposing the variance of observed household expenditure per capita. There are several ways to do this (see, for instance, Combes, Duranton, and Gobillon 2008). A relatively straightforward approach is to algebraically decompose the total variance into four components: 1) the variance of the returns on household characteristics, 2) the variance of location effects, 3) twice the covariance between returns to household characteristics and location effects, and 4) the variance of the residuals.

The contribution of these four components to the overall variation in household expenditure per capita changes quite substantially as locations are introduced in the regression and disaggregated with increasing granularity (Table 3). The contribution of household characteristics falls from 51.5 percent of the total variance in a model without location effects to 39.5 percent in our benchmark specification. In parallel, the

explanatory power of location effects increases from 0 to 12.9 percent, whereas the contribution of the interaction term increases from 0 to 9.8 percent.

As a robustness check, we also decompose the explained variance under the various specifications following a framework based on the Shapley-value function (Huettner and Sunder 2012, and Shorrocks 2013). This methodology allocates the explained variance to the individual explanatory variables based on their marginal contributions. The results, available on request, confirm the growing importance of location as the spatial granularity of the analysis increases.

Locations effects not only attenuate the contribution of household characteristics: they also lead to statistically different estimates of their effects. To illustrate this point we classify household characteristics into four groups: demographics, skills, assets and housing, and social and religious background. For each group of characteristics, we compare the coefficients estimated with our benchmark specification to the coefficients estimated in the model without location effects. Chi-square tests confirm that the estimates are significantly different for all four groups of characteristics. The difference remains statistically significant when comparing with the other two models, in which location effects are defined at state and district levels. This results suggest that conducting poverty or spatial analyses at the state or district levels yields biased results, and that further spatial disaggregation is required to analyze the rural-urban transformation.

Introducing location effects at a disaggregated level also changes the interpretation of the contribution some household characteristics make to household expenditures per capita. We plot the absolute values of the estimated coefficients on household characteristics from our benchmark specification against the estimates from the model without location effects. For educational attainment, we report the marginal effect at the average years of schooling instead of the estimated coefficients. We do so because the square of education variable is also entered in the benchmark specification, to account for possible non-linearity in returns to skills. The estimated coefficients (and marginal effect) decline for 24 of the 35 household characteristics considered, and remain unchanged for only three of them (Figure 4).

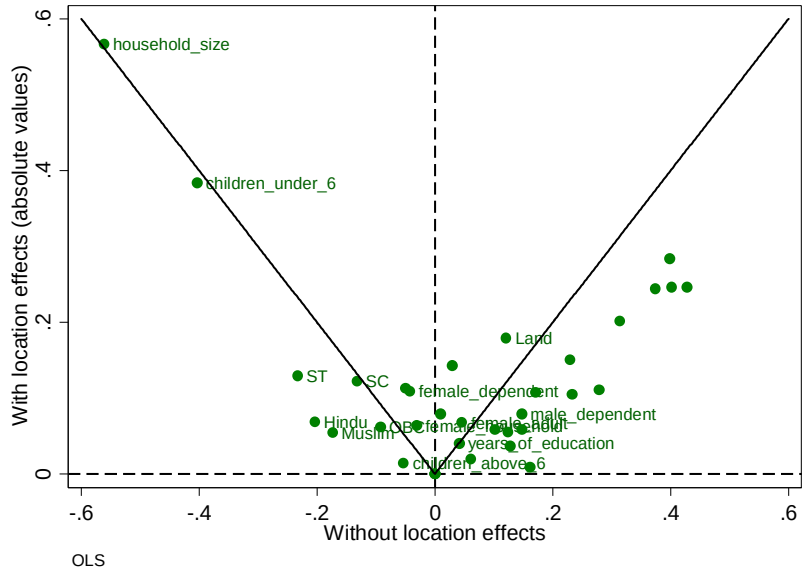
One of the most dramatic changes concerns the estimated effects of the household's social and religious background. Belonging to a Scheduled Tribe, a Scheduled Caste or Other Backward Castes has traditionally been associated with enjoying lower household expenditure per capita. Hindu, and especially Muslim households are also seen as faring worse than Christian households. However, when using our benchmark specification these "stigma" effects decline substantially, as shown in Figure 5. The estimated coefficient on being Hindu falls (in absolute terms) from -0.204 to -0.069, and the coefficient on being Muslim falls (in absolute terms) from -0.173 to -0.054. The drop is similar for the coefficient associated with belonging to Scheduled Tribes, which falls (in absolute terms) from -0.233 to -0.130. The estimated coefficients remain all statistically significant, but it is legitimate to wonder whether ever greater spatial granularity in the analyses would not make them fade away altogether.

6. Cities and catchment areas

The estimated location effects provide a useful metric to evaluate the performance of difference places across India. To make this metric more intuitive, we rescale the location effects by subtracting the median across all 1,406 places. The distributions of location effects are quite spread out. The value of the rescaled location effects ranges from -0.693 to 0.839, with their distribution centered at zero by construction.

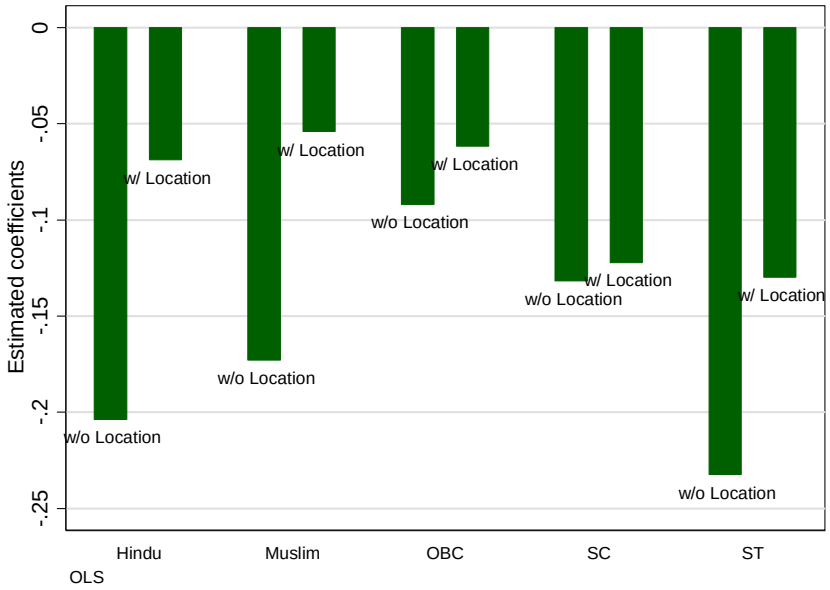
Because household expenditures per capita are measured in log, these figures should not be interpreted as percentages. But they can be converted easily, and they imply that households with the same characteristics have expenditures per capita which are on average 131 percent above the median in the top locations, and 50 percent below the median in bottom locations.

Figure 4. Coefficients on household characteristics



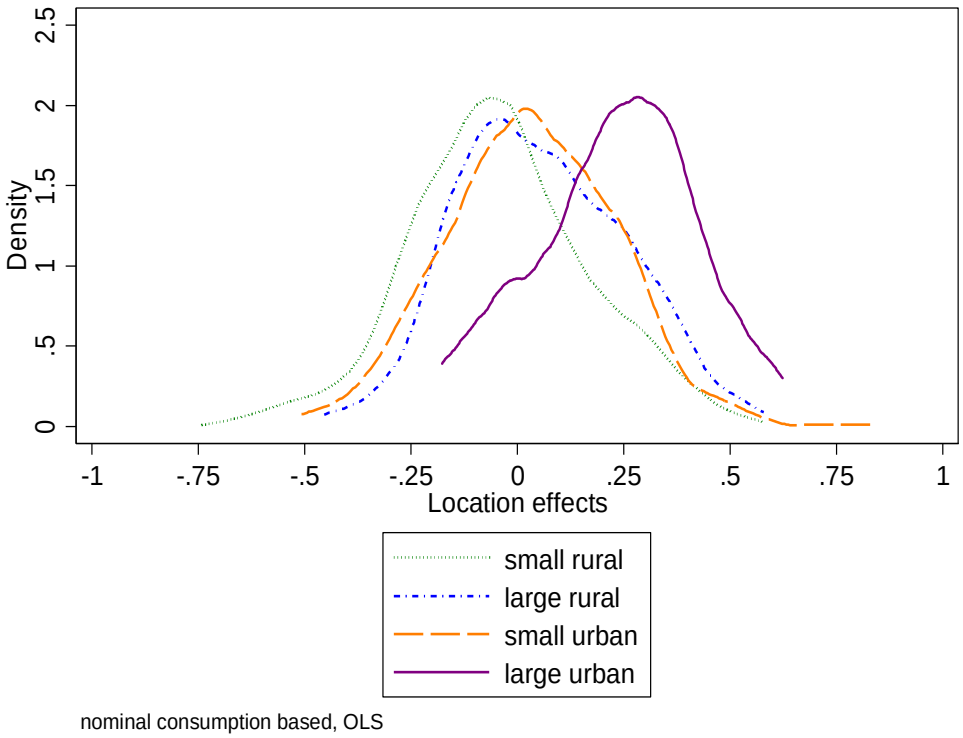
Note: Coefficients on the left of the vertical dotted line are negative, while those on the right are positive. The solid line has a 45-degree slope, corresponding to the case where coefficients are identical.

Figure 5. Coefficients on social and religious backgrounds



Unsurprisingly, location effects are generally higher in urban areas than in rural areas and in places with a larger population than otherwise, as shown in Figure 6. T-tests show that the mean of location effects for small rural areas is smaller than that for large rural areas, and the mean of location effects for small urban areas is smaller than that for large urban areas. Kolmogorov-Smirnov tests further confirm that the distribution of the location effects for the two rural groups differ significantly, as also do the distributions of location effects for the two urban groups.

Figure 6 Distribution of location effects, by population size groups



Note: Location effects are measured relative to the median place in India, and expressed in log. The percent equivalent of a location effect $\gamma_l *$ expressed in log relative to the median place is $100 \cdot [exp(\gamma_l *) - 1]$

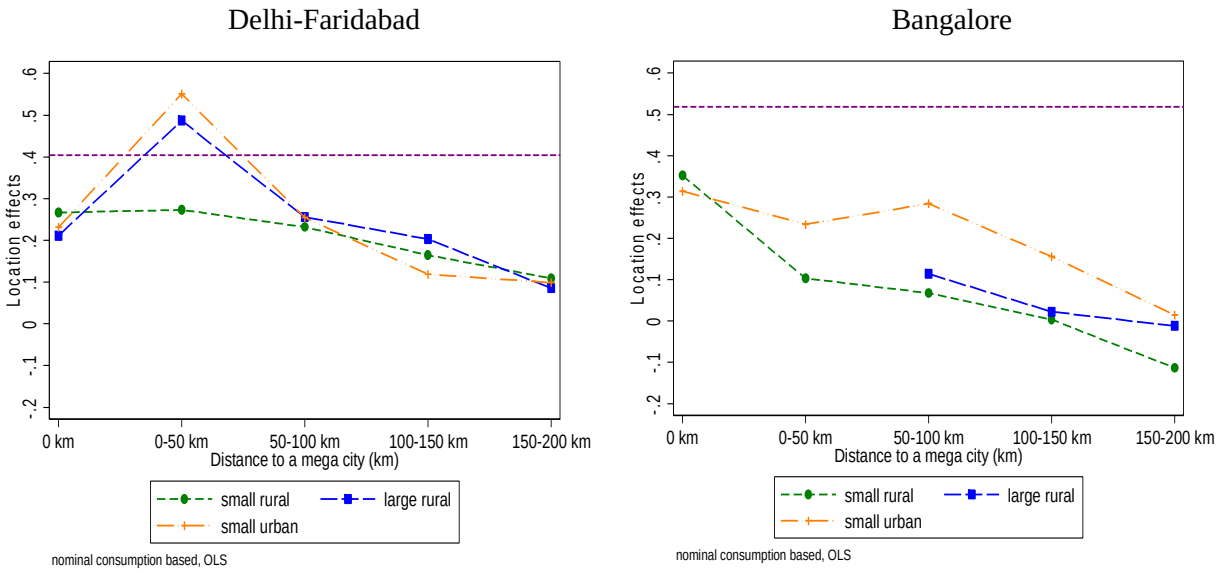
However, the ranking of places is not as straightforward as the notion of a rural-urban divide would suggest. The four distributions have a wide common support, with location effects being sizeable in some rural areas, and clearly below the median in some urban areas. The notion of a rural-urban divide is further undermined by the fact that the distributions of location effects for large rural areas and for small urban areas are difficult to distinguish from each other. T-test cannot reject that the means of the location effects for the two groups are the same. Kolmogorov-Smirnov test cannot reject that the distributions of the location effects for the two groups are the same either. Thus, consistent with the findings by Chatterjee et al. (2015), India seems to be characterized by a rural-urban gradation more than by a rural-urban divide.

Location effects depend not only on the places themselves: they are also influenced by their neighborhoods. As shown before, there is a strong spatial correlation between location effects, even across places belonging

to different population size groups. This suggests that distance, and especially "distance to what?" matters. Because of this high spatial correlation, places near solid performers can be expected to perform well. This is in line with the idea of clustering, and of productive spillovers from the core of the cluster to its periphery. It is also consistent with evidence from advanced economies where the effect of agglomeration economies attenuates with distance (Rosenthal and Strange 2004 and 2008, and Melo et al. 2009). Bottom locations tend to cluster as well, suggesting that improving their performance might be difficult, because that requires countering bad-neighborhood effects.

The importance of "distance to what?" can be illustrated by comparing the places surrounding Delhi and Faridabad to those surrounding Bangalore (Figure 7). Both urban agglomerations are among India's best performers, although Bangalore (with a location effect of 0.512) arguably does better than Delhi and Faridabad (0.415 on average). However, the places surrounding Delhi and Faridabad register much higher location effects on average than those surrounding Bangalore. In fact, location effects for small urban and large rural places within 50 km of Delhi are on average stronger than the average location effect of Delhi and Faridabad. The spread of places with sizeable location effects is also much broader around Delhi and Faridabad than around Bangalore, exceeding 0.1 (a 10.5 percent premium) up to 200 km away from the core. In contrast, the location effects of small rural and large rural places surrounding Bangalore fall below 0.1 after 100 km. This comparison suggests that Bangalore is more productive than Delhi and Faridabad, but its periphery is less productive than the periphery of Delhi and Faridabad.

Figure 7. Delhi-Faridabad versus Bangalore



Note: Location effects are measured relative to the median place in India, and expressed in log. The dotted purple line represents the (average) location effects of the central cities. The other lines represent the average location effects of places surrounding the cities by 50km rings of distance.

Building on the insights from this comparison, we classify all 1,406 places into four tiers. We do so based on both their location effects and their neighborhoods, but ignoring their administrative classification as

urban or rural. The four tiers considered are: 1) top locations, 2) catchment areas, 3) average places, and 4) bottom locations. A cluster is made of one or several top locations plus the associated catchment areas. There is some similarity between our approach to identify clusters and the approaches used to define functional urban areas in advanced economies (OECD 2013, and US Bureau of Census 2011). But our approach is arguably stricter in that it uses the estimated location effects, rather than population density, as the key indicator to generate the classification of places.

We start by identifying *top locations*, defined as the 100 places with the largest location effects. There is some arbitrariness in choosing the number 100 (about 7 percent of all places). But an advantage of this choice is that it facilitates comparisons with urban rankings generated in connection with ongoing government programs in India.

As a second step we look at all other places whose location effect is greater than one standard deviation above the mean, a threshold for better than average, with the goal of identifying which ones belong to a *cluster*. For this, we use a recursive approach. Initially each top location is treated as a cluster, but the cluster is sequentially enlarged to encompass all places above the threshold in the same or contiguous districts. We repeat the process until there are no more places with a location effect a standard deviation above the mean whose districts are contiguous to those in the cluster. It is clear that large clusters such as Delhi-Faridabad can include several top locations, so that there are much fewer clusters than there are top locations. As for the set of places in a cluster which are not top locations, in what follows we refer to them as *catchment areas*.

At the other end, we call *bottom locations* the places (218 in all) whose location effect is more than one standard deviation below the mean. Again, there is some arbitrariness in the chosen threshold, but it allows us to capture a sizeable population mass in this tier. We call *average locations* all the places that do not belong to a cluster and do not quality as bottom locations.

The average location effect for top locations is 0.409, whereas the average for bottom locations is -0.305, a difference that is statistically significant. Interpreted literally, this difference would mean that an average Indian household moving from a bottom location to a top location could expect to see its nominal household expenditure per capita more than double. The gap is obviously wider when considering the extremes of the distribution. Continuing with the hypothetical example, an average Indian household moving from a small rural area in the Malkangiri district of Orissa (the lowest location effect) to Gurgaon city in Haryana (the highest) would see its nominal household expenditure per capita increase 3.6 times.

Remarkably, among the top 100 locations there are more small urban areas (39 in all) than large urban areas (only seven). The ten small urban areas with the largest location effects are: Gurgaon in Haryana, Thiruvananthapuram, Idukki and Kottayam in Kerala, Raigarh in Mahrashtra, Gautam Buddha Nagar and Ghaziabad in Uttar Pradesh, Kachchh in Gujarat, Papum Pare in Arunachal Pradesh and Dakshina Kannada in Karnataka. The seven large urban areas that qualify as top locations are (in descending order) Mumbai in Maharashtra, Bangalore in Karnataka, Faridabad in Haryana, Thane in Maharashtra, Kolkata in West Bengal, Surat in Gujarat, and Delhi. Other large urban areas, such as Agra, Kanpur Nagar and Varanasi in Uttar Pradesh and Patna in Bihar, have location effects below the Indian average for all places, urban and rural. Even more remarkably, about half of India's top locations are administratively rural. This suggests that location, and especially "distance to what?" matter more than administrative status.

The conclusion is similar when focusing on bottom locations. A large majority of them are rural, and 129 of them are actually small rural areas. But there are also 73 small urban areas in this group. Location, again,

seems to be quite decisive. A vast majority of the bottom locations concentrate in the middle of India, crossing the states of Madhya Pradesh, Chhattisgarh, and Orissa from West to East. A number of bottom locations can also be found in Uttar Pradesh and Bihar, along the Ganga basin. Interestingly, most of them are not rural areas but rather small urban areas with exceptionally low location effects.

The importance of location is highlighted by the spatial distribution of clusters (Map 1). The cutoff points used in our approach lead to the identification of 17 clusters in India's case. Geographically, these clusters can be found in the northwest towards Pakistan, in the coastal areas of the mid-west and the southwest, in some inland areas and coastal areas of the southeast, and in the northeast towards Bangladesh and China. Some of these clusters cover multiple districts and, in some cases, multiple states. For example, the cluster of Delhi, Faridabad and Gurgaon spreads across 60 districts in seven northwestern states and union territories. Similarly, the cluster of Thiruvananthapuram includes 19 districts in Karnataka, Kerala and Tamil Nadu. The cluster of Mumbai, Surat and Thane encompasses nine districts in Gujarat and Maharashtra; the cluster of Ahmadabad covers seven districts in Gujarat; and the cluster of Bangalore includes five districts in Karnataka and Tamil Nadu.

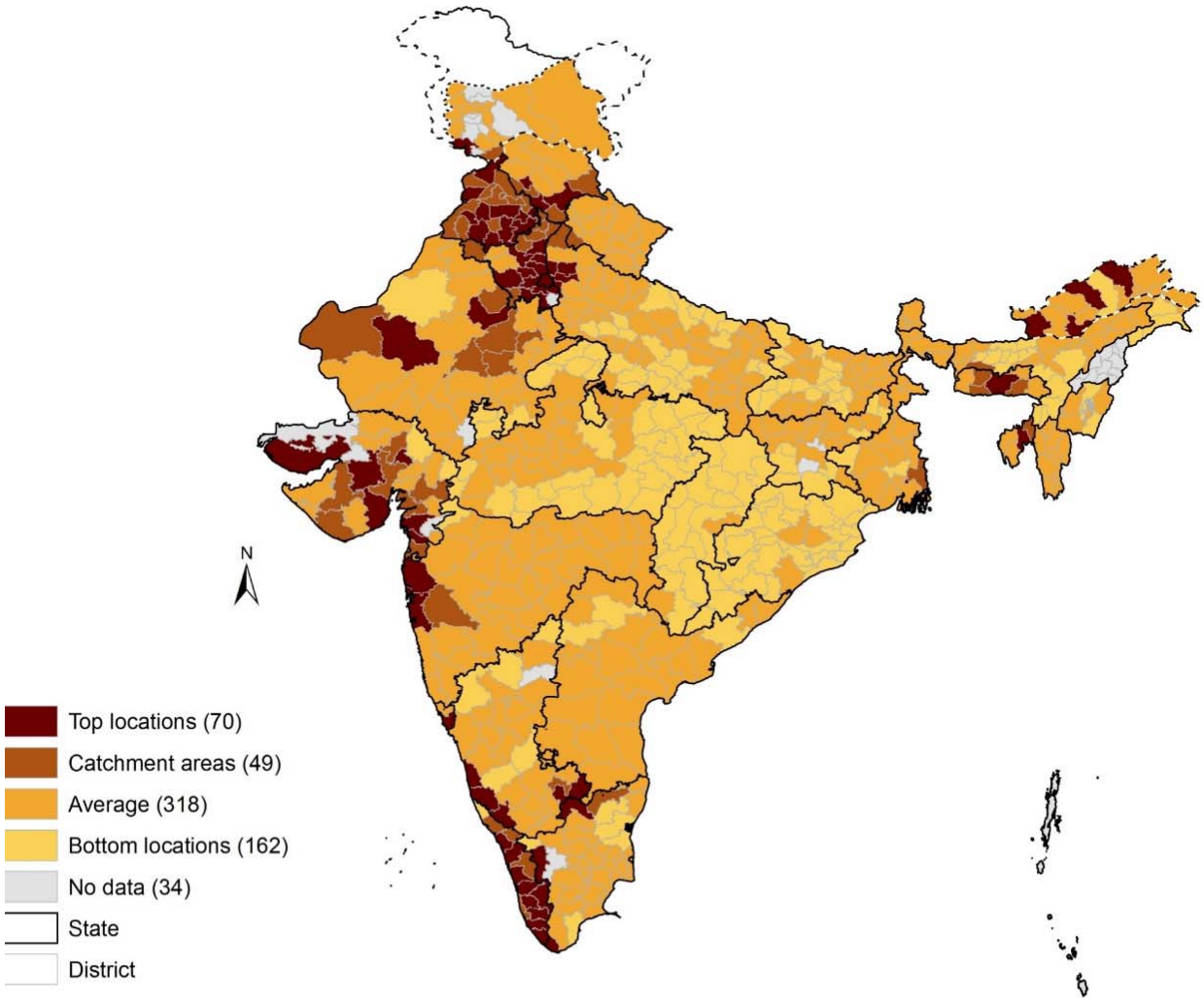
The distance indicator introduced above can be used to assess the geographic spread of a cluster. Consider a cluster whose core is in one district but whose catchment area includes several neighboring districts. The geographic spread can be measured as the maximum distance to the core from all the neighboring districts belonging to the catchment area. By this criterion, the biggest three clusters in India are those of Jodhpur, of Delhi, Faridabad and Gurgaon, and of Ahmadabad, in that order. Another way to measure the size of a cluster is based on its total population. By this metric, the largest three clusters are the one of Delhi, Faridabad and Gurgaon, the one of Thiruvananthapuram and the one of Mumbai, Surat and Thane.

By construction, clusters can include places belonging to all population size groups. Within the 17 clusters identified based on our approach there 12 large urban areas, 91 small urban areas, 45 large rural areas, and 67 small rural areas. Unfortunately, these places can only be mapped at the district level because the NSS 2011-12 does not provide information on the exact geographic position of the different population size groups within a district (Map 2). In principle, a district can include top locations, catchment areas, average locations and bottom locations. Such level of heterogeneity is uncommon given the high spatial correlation of location effects. But it can be found in practice. One salient example is the district of Thane of Maharashtra, part of the Mumbai, Surat and Thane cluster. In Thane, urban areas are among top locations, whereas small rural areas belong to bottom locations. Without going to such extremes, 34 percent of districts include places belonging to two or more different tiers.

7. Location and social inclusion

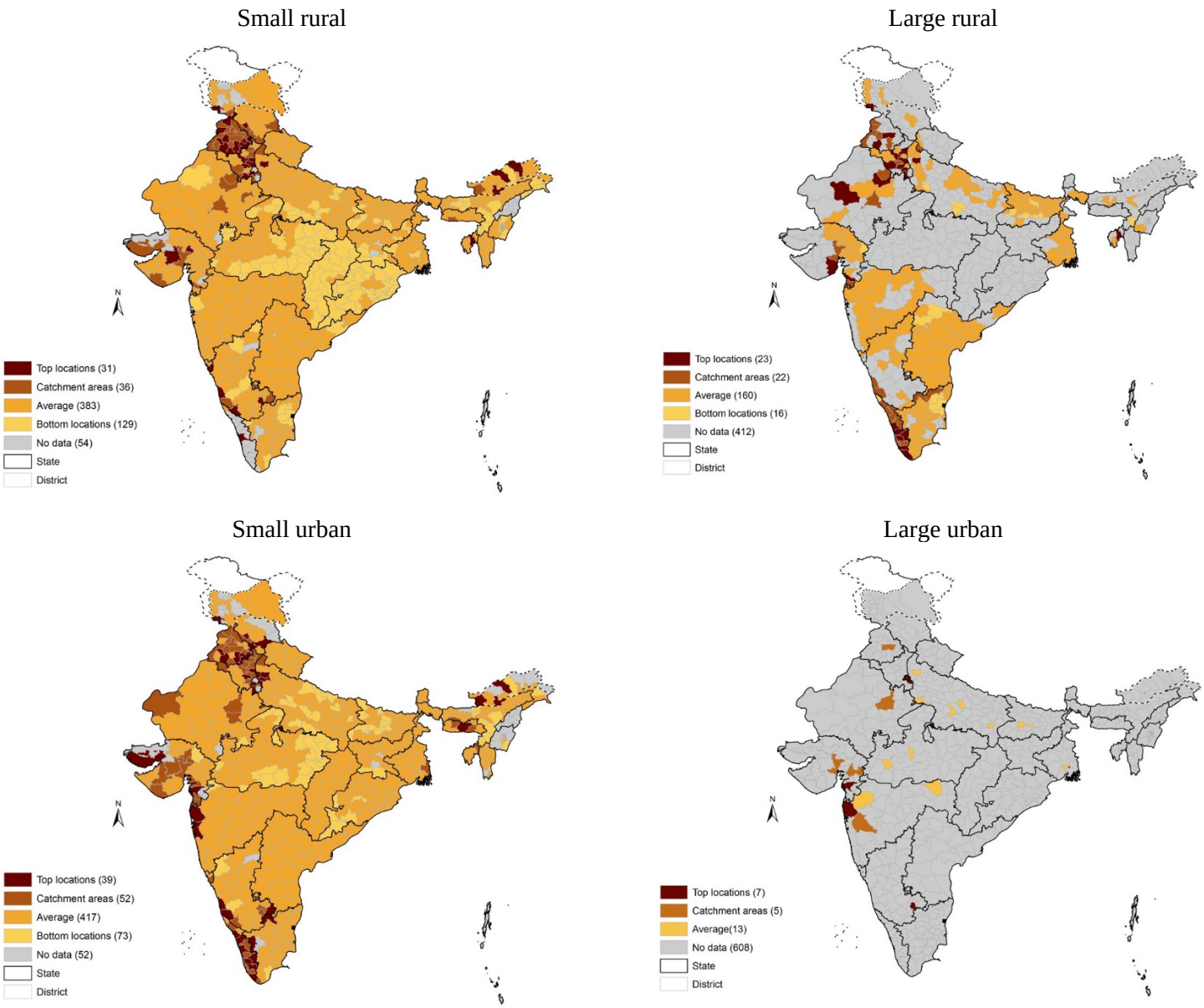
Assessing the relationship between location and social inclusion is not straightforward. To begin with, our location effects are estimated based on nominal household expenditure per capita, which can be seen as a proxy for labor productivity but should not be confused with a proxy for living standards. Indeed, places with higher productivity can be expected to also be characterized by higher land rents, and these in turn may create an upward pressure on the prices of non-tradable goods and services. Households living in places with high location effects may therefore enjoy higher nominal expenditure per capita, but they are also confronted with a higher cost of living.

Map 1. Geographical distribution of the four tiers of locations



Note: Figures in parentheses indicate the number of districts falling into each tier, based on the highest location effect at below-district level.

Map 2. The four tiers of locations by population size group



Note: The figures in parentheses indicates the number of places in the corresponding population size range that fall into each tier.

Price disparities are not a concern when assessing social inclusion within the same place. Indeed, it is safe to assume that households living within a relatively narrow geographic area face roughly the same cost of living. In that case, an inequality measure based on nominal household expenditure per capita is equivalent to an inequality measure based on real household expenditure per capita.

On the other hand, conducting comparisons across places ideally requires correcting for differences in local prices, and this is unfortunately difficult. The price deflators used for poverty analysis in India lack the necessary granularity because they are computed at the state level, only distinguishing between urban and rural areas. An alternative would be to estimate the average rent paid by households in different places. But the NSS does not report imputed rent for households that own or occupy their dwelling, and information on housing characteristics is anyway too sparse to generate comparable rent measures across space. When comparing households across places we therefore have no other choice than to assume that nominal and real household expenditure per capita are highly correlated. This is a plausible assumption, but differences in the availability of affordable housing across places make the correlation less than perfect.

With this measurement caveat in mind, location effects provide a first and useful approximation to inclusion. The notion of a rural-urban divide suggests that urbanization ought to be associated with an increase in inequality, at least until a substantive majority of the population lives in urban areas. But this presumption may not be correct in the presence of a rural-urban gradation, especially when high-productivity clusters encompass large numbers of administratively rural areas, as is the case in India. Based on our classification of places into four tiers, 31.8 percent of the population in top locations is administratively rural, and the proportion increases to 51.4 percent in their catchment areas (Table 4). In fact, more than 10 percent of the rural population live in one of India's 17 clusters, compared to about 18 percent in its 218 bottom locations.

Moreover, the location effects of the rural places encompassed by the 17 clusters are not very different from the location effects of the urban places in them. Among top locations, the average location effect of rural places is 0.40, not far below the 0.46 average for urban places. The gap is even smaller in catchment areas (0.27 versus 0.28 respectively). From this perspective, large spread-out clusters such as the one around Delhi, Faridabad and Gurgaon, or that around Thiruvananthapuram, make a positive contribution to social inclusion. Rather than exacerbating the alleged rural-urban divide, they make the surrounding rural populations benefit from the dynamism of their cores.

A different issue is whether inequality within a given place increases as this place becomes more productive. The presumption is that it does: stronger agglomeration economies often rests on a subset of the population being highly skilled, and rapid urban growth may also be associated with temporary rents in leading sectors of activity. To assess whether this presumption is correct, we compute the mean log deviation (MLD) of household expenditure per capita in each of the 1,406 places considered in the analysis. We do the same for each of the clusters comprising a large urban area, using the NSS weights to that effect. The MLD is a standard measure of inequality. It can be interpreted as the gap in expenditure per capita between a randomly selected person in a given population and the average person in the same group. The greater the extent of local inequality, the larger the gap. In comparison with other inequality indicators, MLD attaches equal weight to each observation and thus captures the level of inequality across the full range of households.

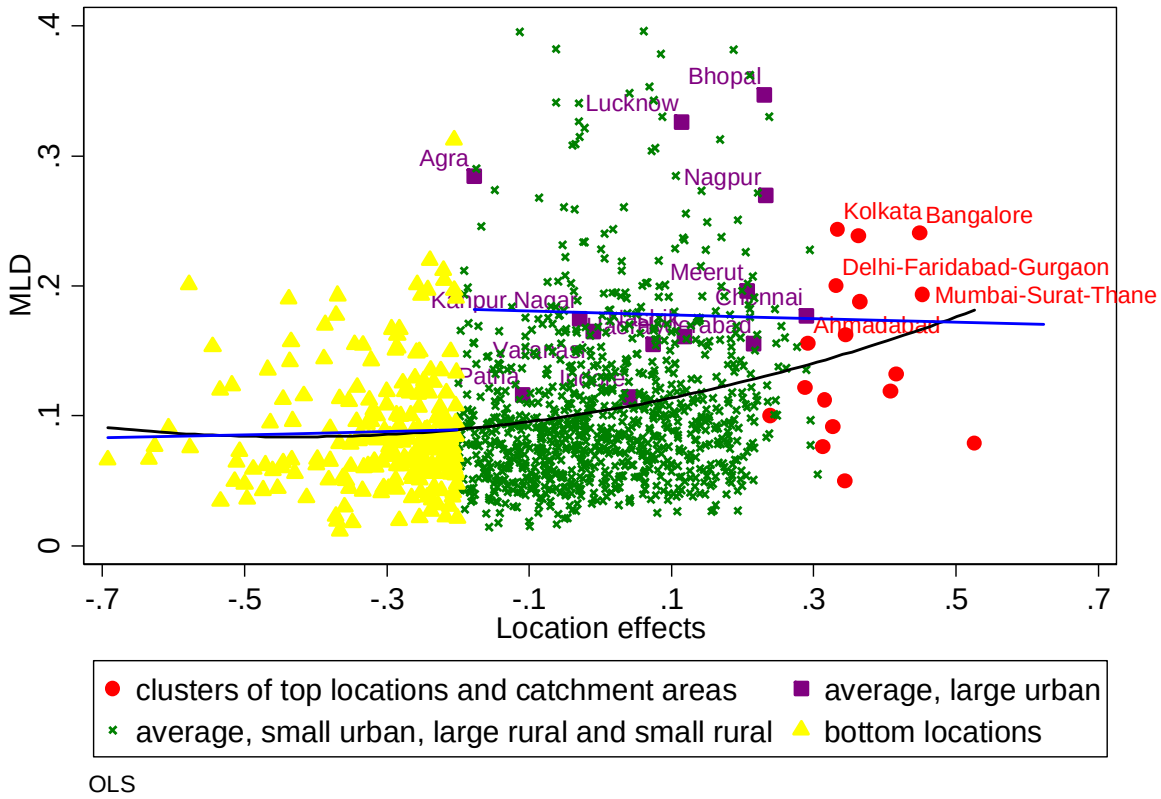
Table 4. Population shares and location effects by rural and urban and by four tiers of locations

Population by tier (percent)							
	Small rural	Large rural	Small urban	Large urban	Rural	Urban	Total
Clusters	6.9	21.8	24.4	65.1	10.1	36.7	17.7
Top locations	2.6	10.4	12.4	47.3	4.3	23.0	9.6
Catchment areas	4.3	11.4	12.0	17.8	5.8	13.7	8.1
Average locations	72.0	72.0	68.8	34.9	72.0	58.5	68.1
Bottom locations	21.2	6.2	6.9	0.0	17.9	4.8	14.2
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Population by type of place (percent)							
	Small rural	Large rural	Small urban	Large urban	Rural	Urban	Total
Clusters	21.6	19.1	27.4	31.9	40.7	59.3	100.0
Top locations	15.0	16.8	25.6	42.6	31.8	68.2	100.0
Catchment areas	29.5	21.9	29.5	19.1	51.4	48.6	100.0
Average locations	59.0	16.4	20.1	4.4	75.5	24.5	100.0
Bottom locations	83.5	6.8	9.7	0.0	90.4	9.7	100.0
Total	55.9	15.6	19.9	8.7	71.4	28.6	100.0
Average location effect (weighted by population)							
	Small rural	Large rural	Small urban	Large urban	Rural	Urban	Total
Clusters	0.31	0.35	0.35	0.42	0.33	0.39	0.37
Top locations	0.39	0.42	0.44	0.47	0.40	0.46	0.44
Catchment areas	0.26	0.28	0.27	0.29	0.27	0.28	0.27
Average locations	-0.03	0.01	0.04	0.13	-0.02	0.06	0.00
Bottom locations	-0.30	-0.27	-0.27		-0.30	-0.27	-0.30

Note: Location effects are measured relative to the median place in India, and expressed in log.

The relationship between MLD and location effects is convex (Figure 8). At the low end, in places where labor productivity is low, the relationship is actually flat: increases in the location effect are not associated with increases with the MLD. As expected, a positive relationship emerges when moving to locations with higher labor productivity. But there are also important nuances at the top end, as shown by the straight line in the upper right side of the figure. This line represents the estimated relationship between location effects and MLD for all large cities in India (both clusters encompassing large urban areas and other large urban areas). The line is slightly downward-sloping, but the slope is not statistically significant. This means that the most productive urban centers in India are not more unequal than other major cities. This result is partly driven by the relatively poor performance of large urban areas in the Ganga basin. Thus, Agra registers much lower productivity and much higher inequality than the cluster of Delhi, Faridabad and Gurgaon. And Kanpur and Varanasi show much lower productivity than the cluster of Mumbai, Surat and Thane for a similar level of inequality.

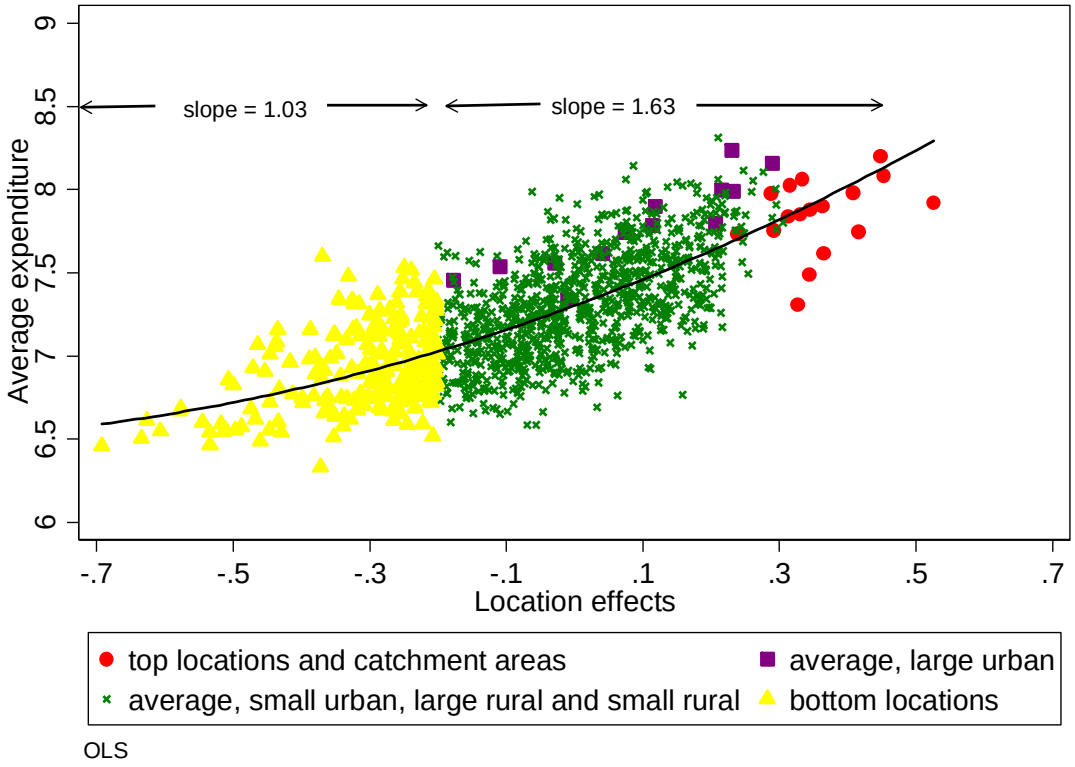
Figure 8. Location effects and inequality



Note: Location effects are measured relative to the median place in India, and expressed in log. The convex line represents the fitted quadratic relationship across all clusters and other locations. The straight line on the upper right side represents the fitted linear line for large cities, including clusters with large urban areas as well as large urban areas belonging to average locations. The straight line on the lower left side represents the fitted linear line for bottom locations.

Household sorting is another mechanism through which larger location effects can result in greater inequality. The comparisons above are based on the hypothetical example of an average household moving across different places in India. But household characteristics vary across locations. Average expenditure per capita increases with location effects, as could be expected. But it increases more than proportionally (Figure 9). The slope of the relationship is close to one when location effects are small, meaning that a 10 percent increase in location effects is associated with an increase in average expenditure per capita by roughly 10 percent. But the relationship is convex, implying that the associated increase in average expenditure per capita becomes greater as location effects increase. Such pattern is indicative of sorting, with households whose characteristics are associated with greater returns living in more productive places. Household sorting amplifies the spatial gaps in living standards, beyond what the dispersion in location effects would suggest.

Figure 9. Location effects and average expenditure



Note: Location effects are measured relative to the median place in India, and expressed in log. Average expenditures are measured in log. The convex line represents the fitted quadratic relationship across all clusters and other locations. The slope of the relationship is indicated at the top of the figure, distinguishing between bottom locations and the rest.

The contribution sorting makes to inequality can be assessed by decomposing the variance of observed household expenditure across the four tiers of locations. The total variance is the sum of the variance explained by the benchmark specification and the variance associated with the residuals. The former component can be expressed as the sum of the variance within each of the tiers, the variance between tiers, and a cross term. Both the variance within tiers and the variance between tiers can be further decomposed into three sub-components: 1) the variance of the returns on household characteristics, 2) the variance of location effects, and 3) twice the covariance between returns to household characteristics and location effects. The algebraic decomposition is presented in Annex 1.

Variance in household characteristics is the main contributor to the variance in expenditure per capita within a specific tier (Table 5). This is especially so for the relatively large group of average locations. On the other hand, location effects and the interaction term between them and household characteristics account for much of the variance between tiers. The two together contribute 16.7 (= 9.9 +6.8) percentage points to the 17.9 percent of the total variance in household expenditure accounted for by variance between tiers. This pattern is the most significant for the top and the bottom locations. Because the interactions captures sorting by households, their large powers in explaining between tier variations confirm the significance of

sorting in amplifying inequality in household expenditure per capita, particularly at both ends of the distribution.

Table 5. Variance decomposition by four tiers of location

	Bottom locations	Average locations	Catchment areas	Top locations	All
Value					
Observed expenditure	0.060	0.206	0.043	0.079	0.388
Predicted expenditure	0.046	0.122	0.024	0.049	0.241
Within	0.020	0.120	0.014	0.017	0.172
Household	0.018	0.101	0.014	0.016	0.148
Location	0.001	0.009	0.000	0.001	0.012
Interaction	0.000	0.011	0.000	0.001	0.012
Between	0.026	0.001	0.010	0.032	0.070
Household	0.001	0.000	0.001	0.003	0.005
Location	0.016	0.000	0.005	0.016	0.038
Interaction	0.008	0.001	0.004	0.013	0.026
Cross	0.000	0.000	0.000	0.000	0.000
Percentage					
Observed expenditure	15.4	53.1	11.2	20.3	100.0
Predicted expenditure	11.8	31.4	6.2	12.8	62.2
Within	5.1	31.1	3.6	4.5	44.2
Household	4.7	25.9	3.6	4.0	38.3
Location	0.3	2.3	0.0	0.3	3.0
Interaction	0.0	2.8	0.0	0.2	3.0
Between	6.7	0.4	2.7	8.3	17.9
Household	0.3	0.1	0.2	0.7	1.3
Location	4.2	0.1	1.4	4.1	9.9
Interaction	2.2	0.2	1.0	3.4	6.8
Cross	0.0	0.0	0.0	0.0	0.0

8. Correlates of location effects

An obvious question, in light of the wide spatial disparities in location effects and household expenditure per capita, is what makes for good places. In particular, some clusters and top locations are characterized by high labor productivity and not necessarily by higher inequality. From a policy making perspective, it would be important to understand which location characteristics support such superior performance. This understanding would help policy makers maximize the payoffs from the rural-urban transformation.

Unfortunately, the answers are limited so far. Multiple explanations have been proposed in the urban economics literature on the potential sources of agglomeration economies: pooling of labor, sharing of resources and productive amenities, knowledge spillovers and so on (Marshall 1890, Jacobs 1969, Krugman 1991). Other factors that have been listed as potentially stimulating the formation and expansion of cities are pleasant climate and abundance of natural resources and so on (Glaeser and Gottlieb 2009). But in the end, this is an empirical question and the answer is bound to vary from country to country.

Carefully exploring the role of these potential channels and factors in India's case is a research agenda on its own. And solid answers are bound to require not just the estimation of a cross-section of location effects, as we have done in this paper. A dynamic analysis on how the magnitude of the location effects evolves over time is also required. This may be the most promising way of identifying the best predictors of strong local performance, and our research following up on this paper goes in that direction. For now, as a preliminary probe into that direction, we assess the correlations between the factors that potentially drive agglomeration economies, or contribute to faster productivity growth, and the tier that a location falls into.

The factors we consider in this correlation analysis are not meant to be exhaustive, but they do capture some of the key channels and factors discussed in the literature. We define a series of spatial indicators related to population density, employment structure, skills, or infrastructure, and we compare the value of these indicators across tiers of locations. The unit of observation is either place or district, depending on data availability. Our reference for the comparison is the value of these indicators for the 100 top locations in the country. We assess whether the values are different across tiers using two-sample t tests for means and two-sample Kolmogorove-Smirnov tests for distributional dominance (Tables 6 and 7 respectively). The detailed location characteristics of each of the clusters, top locations, catchment areas, and other major Indian cities are available on request.

Population density is the first factor of interest. The urban economics literature suggests greater population density is associated with higher productivity driven by agglomeration forces. Consistent with the literature, population density increases from bottom to average locations, from average locations to catchment areas and top locations. Statistically, top locations register significantly higher population density than the other tiers, including catchment areas though the difference is smaller.

A related factor is employment structure. Ultimately, agglomeration economies are driven by firms and workers collocating and engaging more frequently and intensely. Labor pooling is an important channel underlying agglomeration economies. Employment density thus may capture the underlying ideas of agglomeration economies more accurately than population density. In the context of advanced economies, employment density was first introduced by Ciccone and Hall (1996) and has been widely used by the urban economics literature. In the context of developing countries such as India, most people work, especially the poor because they cannot afford not to. The nature of economic activity of the working poor, however, is often different from the nature of general employment discussed in the context of the advanced economies (Banerjee and Duflo 2011, World Bank 2012). Employment density may lose its advantage over population density.

We therefore use employment structure instead, measured through the shares of main workers and regular wage workers in total employment. These two types of jobs are more likely to resemble the jobs found in advanced economies. According to the Census of India 2011, main workers are those who worked for six months or longer in the past 12 months versus marginal workers working less than six months. Based on NSS 2011-12, regular wage workers are those who worked in other people's farm or non-farm enterprises

and got in return a salary or wage on a regular basis. Jobs covered by these two concepts are more likely to resemble the jobs considered in the context of the advanced economies and allow us to better approximate the idea of employment density. In line with the literature, we find that the employment shares of both main workers and regular wage workers increase from bottom locations to top locations. Top locations and catchment areas have statistically higher employment shares for both types of workers than bottom and average locations. The results reported on regular wage workers are based on usual employment status, which includes both usual principal and subsidiary employment status in the past year, as defined by NSS 2011-12, but they are robust to using principal employment status instead.

We also look at the sectoral composition of local economies. Manufacturing and services are generally believed to hold higher potentials for productivity growth than agriculture. We use the employment shares of agriculture and manufacturing as indicators of the structure of economic activity at the local level. Our analysis shows that top locations and catchment areas have both significantly lower employment shares in agriculture and significantly higher shares in manufacture.

Skills have also been consistently found to drive agglomeration economies, especially by facilitating knowledge spillovers. In the context of advanced economies, skills density is often measured by the share of the population with tertiary education. While educational attainment has been improving rapidly in India, it is still much lower than in advanced economies. We thus measure average skills through the share of the working-age population with secondary education, in addition to the share with tertiary education. We find top locations and catchment areas have significantly higher shares of working-age people in these two groups than bottom and average locations.

The opportunity to share productive amenities is another potentially important factor driving agglomeration economies. Following the literature, we use measures on access to infrastructure and public services to capture the idea of productive amenities. Our results show that road density increases from bottom locations to top locations and top locations register significantly higher road density than any other tier. In addition, access to electricity, access to cellphone and access to banking services are all significantly higher in top locations and catchment areas than in other locations.

Finally, we also consider the differences between tiers of locations in terms of the social and religious backgrounds of their populations. In the literature on governance, ethno-linguistic fractionalization has been found to affect the quality of institutions and long-run growth. Culture is also increasingly highlighted as a determinant of individual aspirations and social interactions by the literature on behavioral economics. Discrimination toward people of lower castes, tribal groups and religious minorities has been a long-standing concern in India as well. Our results suggest that Scheduled Tribes are much more concentrated in bottom locations than in any other tiers of locations. Scheduled Castes, on the other hand, do not exhibit a greater concentration in bottom and average locations than in top locations and catchment areas.

Overall, catchment areas resemble top locations, except in their population density and road density. Top locations and catchment areas register more main workers and more regular wage workers, lower agriculture employment and more employment in manufacturing, higher educational attainment at both secondary and tertiary levels, and better access to infrastructure and services. These results are consistent with the higher labor productivity observed in top locations and catchment areas, relative to the rest of India. Bottom locations stand out as places with higher concentration of Scheduled Tribes. However, in interpreting these results it is important to keep in mind that correlation is not causality.

Table 6. Average location characteristics by tier

	Bottom locations	Average locations	Catchment areas	Top locations	Unit of observation
Population density					
Population density (people/sq. km, in log)	5.761***	5.831***	6.007*	6.358	District
Type of jobs					
Main workers (% of total workers)	67.241***	73.000***	79.799	82.098	District
Regular wage workers (% of total workers)	13.266***	21.479***	30.094	32.418	Place
Sectoral structure					
Agriculture share (% of total workers)	47.847***	39.230***	28.375	28.896	Place
Manufacturing share (% of total workers)	7.881***	11.879***	16.995	15.694	Place
Skills					
Secondary education (% total working age population)	25.117***	26.795***	33.885	33.277	Place
Tertiary education (% total working age population)	8.168***	10.170***	12.538	14.368	Place
Infrastructure and services					
Road density (km/ sq. km)	0.163***	0.284***	0.488**	0.751	District
Access to electricity (% of total households)	48.780***	64.151***	85.45	89.87	District
Access to cellphone (% of total households)	44.551***	56.547***	69.889	72.594	District
Access to banking services (% of total households)	51.288***	57.771***	66.775	68.118	District
Social and religious background					
Scheduled Tribes (% of total population)	20.844***	12.182	6.849	9.215	Place
Scheduled Castes (% of total population)	16.197*	16.396**	20.829	19.989	Place

Note: The numbers are means for each tier. The significance of t-test values for the difference with the mean of top locations is indicated by * for the 0.1 level, ** for the 0.05 level, and *** for 0.01 level.

Table 7. Distribution of location characteristics by tier

	Bottom locations		Average locations		Catchment areas		
	H0: Top locations have larger values	**H0: Top locations have smaller values**	**H0: Top locations have larger values**	**H0: Top locations have smaller values**	**H0: Top locations have larger values**	**H0: Top locations have smaller values**	**Unit of observation**
Population density							
Population density (people/sq. km, in log)	0.0367	-0.339***	0.0146	-0.2843***	0.0755	-0.2959***	District
Type of jobs							
Main workers (% of total workers)	0	-0.5681***	0	-0.3889***	0.102	-0.149	District
Regular wage workers (% of total workers)	0	-0.5143***	0.001	-0.2788***	0.0591	-0.114	Place
Sectoral structure							
Agriculture share (% of total workers)	0.3401***	-0.0088	0.1966***	0	0.0778	-0.0561	Place
Manufacturing share (% of total workers)	0.0058	-0.3722***	0.0163	-0.211***	0.1209	-0.0226	Place
Skills							
Secondary education (% total working age population)	0.0083	-0.4464***	0	-0.3621***	0.0991	-0.0791	Place
Tertiary education (% total working age population)	0	-0.2818***	0	-0.1853***	0.0874	-0.1235	Place
Infrastructure and services							
Road density (km/ sq. km)	0.0101	-0.7051***	0.0128	-0.5121***	0.0857	-0.2224*	District
Access to electricity (% of total households)	0	-0.7317***	0	-0.5419***	0.0143	-0.151	District
Access to cellphone (% of total households)	0	-0.773***	0.008	-0.6619***	0.0143	-0.1918	District
Access to banking services (% of total households)	0.0081	-0.5176***	0.006	-0.3446***	0.0469	-0.0755	District
Social and religious affiliations							
Scheduled Tribes (% total population)	0.3842***	-0.0179	0.2143***	-0.009	0.073	-0.0491	Place
Scheduled Castes (% total population)	0.0239	-0.1312*	0.0053	-0.1074	0.0878	-0.0565	Place

Note: The numbers report the Kolmogorov–Smirnov two one-way test results comparing the distribution of other tiers with that of top locations. The significant levels are reported as well: *significant at 0.1 level, ** significant at 0.05 level, *** significant at 0.01 level.

9. Conclusion

In this paper, we combine the insights of poverty analysis and urban economics and develop a workable metric to describe and understand spatial patterns of labor productivity across India. Relying on this metric, we highlight the importance of location as a determinant of economic performance. The estimated location effects, capturing the labor productivity premium associated with narrowly-defined places, are consistent with the notion of a rural-urban gradation. While urban places perform better on average than rural places, the boundaries between the two are blurred, with large rural areas and small urban areas being indistinguishable in some respects. The results also show that distance matters — places with higher location effects tend to be near each other but some spread their prosperity to a greater extent than others. Building on these observations, we use the distribution of this metric and spatial contiguity to classify places into four tiers: top locations, their catchment areas, average locations and bottom locations. This way, we break the barriers set by administrative boundaries and administrative status and identify 17 clusters of top locations and their catchment areas. Many of these clusters include high performing rural areas. And in spite of their high performance these clusters are not necessarily more unequal than other, less well-performing large urban areas.

Our preliminary analysis also suggests these clusters are associated with factors that potentially drive agglomeration economies and contribute to faster productivity growth. For example, both top locations and catchment areas register more main workers and regular wage workers than average locations and bottom locations. They also have better skills, as reflected in higher educational attainment at both secondary and tertiary levels. And they are characterized by better infrastructure and access to services. However, this preliminary characterization of locations is based on static correlations, so the identified correlations should not be interpreted as implying causality. As a follow-up to these findings, we will extend the framework and analyze inter-temporal changes in performance across places. This will hopefully allow us to understand the emergence of clusters and other high-performing locations, and to assess whether the rural-urban transformation is characterized by convergence or divergence.

References

Acemoglu, Daron, and Joshua Angrist. "How large are human-capital externalities? Evidence from compulsory-schooling laws." *NBER Macroeconomics Annual 2000, Volume 15*. MIT Press, 2001. 9-74.

Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International regional science review*, *26*(2), 153-166.

Anselin, L., & Rey, S. J. (2010). *Perspectives on spatial data analysis* (pp. 1-20). Springer Berlin Heidelberg.

Banerjee, A., Banerjee, A. V., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs.

Bureau of the Census. (2011). Urban Area Criteria for the 2010 Census. Federal Register 76(164). Bureau of the Census, Department of Commerce, Government of the United States.

Chatterjee, U., Murgai & Rama, M. (2015). Job Opportunities along the Rural-Urban Gradation and Female Labor Force Participation in India. *Unpublished manuscript*. Washington DC: The World Bank.

Ciccone, A., & Hall, R. E. (1996). Productivity and the density of economic activity. *The American Economic Review*, *86*(1), 54.

Combes, P. P., Duranton, G., & Gobillon, L. (2008). Spatial wage disparities: Sorting matters!. *Journal of Urban Economics*, *63*(2), 723-742.

Combes, P. P., Duranton, G., Gobillon, L., & Roux, S. (2010). Estimating agglomeration economies with history, geology, and worker effects. In *Agglomeration Economics* (pp. 15-66). University of Chicago Press.

Deaton, A. (1997). *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications.

Demombynes, G., Elbers, C., Lanjouw, J., Lanjouw, P., Mistiaen, J., & Özler, B. (2002). *Producing an improved geographic profile of poverty: methodology and evidence from three developing countries* (No. 2002/39). WIDER Discussion Papers//World Institute for Development Economics (UNU-WIDER).

Duranton, G. (2014). Growing through cities in developing countries. *The World Bank Research Observer*, lku006.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro–level estimation of poverty and inequality. *Econometrica*, *71*(1), 355-364.

Gill, I. S., & Goh, C. C. (2010). Scale economies and cities. *The World Bank Research Observer*, *25*(2), 235-262.

Glaeser, E. L., & Gottlieb, J. D. (2009). *The wealth of cities: Agglomeration economies and spatial equilibrium in the United States* (No. w14806). National Bureau of Economic Research.

Glaeser, E., & Maré, D. (2001). Cities and Skills. *Journal of Labor Economics*, *19*(2), 316-42.

Glaeser, E. L., & Resseger, M. G. (2010). The complementarity between cities and skills*. *Journal of Regional Science*, *50*(1), 221-244.

Hentschel, J., Lanjouw, J. O., Lanjouw, P., & Poggi, J. (2000). Combining census and survey data to trace the spatial dimensions of poverty: A case study of Ecuador. *The World Bank Economic Review*, *14*(1), 147-165.

Huettner, F., & Sunder, M. (2012). Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. *Electronic Journal of Statistics*, *6*, 1239-1250.

Jacobs, J. (1969). *The Economy of Cities*. New York: Random House.

Jalan, J. & Ravallion, M. (2002). Geographic Poverty Traps? A Micro Model of Consumption Growth in Rural China. *Journal of Applied Econometrics*. 17(4), pp. 329-346.

Kanbur, R., & Venables, A. J. (eds.) (2005). *Spatial Inequality and Development*. World Institute for Development Economics Research, Oxford University Press.

Li, Y., Rama, M., Galdo, V. & Pinto, M. F. (2015). *A Spatial Database for South Asia*. Washington DC: The World Bank, forthcoming.

Krugman, P. R. (1991). *Geography and Trade*. Cambridge, MA: Mit Press.

Marshall, A. (1890). *Principles of Economics*. London: Macmillan.

Melo, P. C., Graham, D. J., & Noland, R. B. (2009). A meta-analysis of estimates of urban agglomeration economies. *Regional science and urban Economics*, *39*(3), 332-342.

Mion, G., & Naticchioni, P. (2009). The spatial sorting and matching of skills and firms. *Canadian Journal of Economics/Revue canadienne d'économie*, *42*(1), 28-55.

Moretti, E. (2004a). Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of econometrics*, *121*(1), 175-212.

Moretti, E. (2004b). Human capital externalities in cities. *Handbook of regional and urban economics*, *4*, 2243-2291.

NSSO. (2012). The 68th round of National Sample Survey of India. National Sample Survey Office, Ministry of Statistics and Programme Implementation, Government of India.

OECD. (2013). Definition of Functional Urban Areas (FUA) for the OECD metropolitan database. OECD, Paris.

Office of the Registrar General and Census Commissioner. (2001). Census of India 2001. Ministry of Home Affairs, Government of India. http://censusindia.gov.in/

Office of the Registrar General and Census Commissioner. (2011a). Census of India 2011. Ministry of Home Affairs, Government of India. http://censusindia.gov.in/

Office of the Registrar General and Census Commissioner. (2011b). Census of India 2011: Administrative Atlas of India. Ministry of Home Affairs, Government of India. http://censusindia.gov.in/

Pradhan, Kanhu Charan. 2013. “Unacknowledged Urbanisation. New census towns in India”. *Economic and Political Weekly* 48 (36).

Puga, D. (2010). The Magnitude and Causes of Agglomeration Economies*. *Journal of Regional Science*, *50*(1), 203-219.

Rauch, J. E. (1993). Productivity gains from geographic concentration of human capital: evidence from the cities. *Journal of urban economics*, *34*(3), 380-400.

Ravallion, M. & Jalan, J. (1999). China's Lagging Poor Areas. *American Economic Review*, 89(2): 301-305.

Roback, J. (1982). Wages, rents, and the quality of life. *The Journal of Political Economy*, 1257-1278.

Rosen, S. (1979). Wage-based Indexes of Urban Quality of Life. In P. N. Miezkowski and M. R. Straszheim (eds.), *Current Issues in Urban Economics*. Baltimore, MD: Johns Hopkins University Press, pp. 74–104.

Rosenthal, S. S., & Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. *Handbook of regional and urban economics*, *4*, 2119-2171.

Rosenthal, S. S., & Strange, W. C. (2008). The attenuation of human capital spillovers. *Journal of Urban Economics*, *64*(2), 373-389.

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: a unified framework based on the Shapley value. *Journal of Economic Inequality*, 1-28.

Wheaton, W. C., & Lewis, M. J. (2002). Urban wages and labor market agglomeration. *Journal of Urban Economics*, *51*(3), 542-562.

World Bank (2012). Jobs. *World Development Report*, 2013. Washington DC: The World Bank.

World Bank (2015). Ending Poverty and Sharing Prosperity. *Global Monitoring Report*. Washington DC: The World Bank.

Annex 1. Decomposition of the variance across tiers of places

Our benchmark specification is

$$Ln\left(\begin{matrix}Nominal\ expenditure\\ per\ capita\end{matrix}\right)_{hl} = \alpha + \beta \cdot \left(\begin{matrix}Household\\ characteristics\end{matrix}\right)_{hl} + \gamma_l + \varepsilon_{hl} \quad (3)$$

Denote $\beta \cdot \left(\begin{matrix}Household\\ characteristics\end{matrix}\right)_{hl}$ as $\beta \cdot H_{hl}$ and use k to represents a tier of locations with $k = 1, \dots, 4$, N to represent total number of households and N^k to represent total number of households of tier k, it follows

$$Var\left(ln\left(\begin{matrix}Nominal\ expenditure\\ per\ capita\end{matrix}\right)_{hl}\right) = \underbrace{Var(\alpha + \beta \cdot H_{hl} + \gamma_l)}_{variance\ of\ predicted\ expenditure} + \underbrace{Var(\varepsilon_{hl})}_{variance\ of\ residuals}$$

$$= \underbrace{\sum_{h=1}^{N} \frac{1}{N} E[((\beta \cdot H_{hl} + \gamma_l) - \overline{(\beta \cdot H_{hl} + \gamma_l)})^2]}_{variance\ of\ preidcuted\ expenditure} + \underbrace{Var(\varepsilon_{hl})}_{variance\ of\ residuals}$$

$$= \underbrace{\sum_{k=1}^{4} \frac{N^k}{N} \sum_{h=1}^{N^k} \frac{1}{N^k} E[((\beta \cdot H_{hl} + \gamma_l) - \overline{(\beta \cdot H_{hl} + \gamma_l)})^2]}_{variance\ of\ predicuted\ expenditure} + \underbrace{Var(\varepsilon_{hl})}_{variance\ of\ residuals}$$

$$= \sum_{k=1}^{4} \frac{N^k}{N} \sum_{h=1}^{N^k} \frac{1}{N^k} \{ \overbrace{E\left[\left((\beta \cdot H_{hl} + \gamma_l) - \overline{(\beta \cdot H_{hl} + \gamma_l)}^k\right)^2\right]}^{Sum\ of\ the\ variance\ of\ each\ tier} + \overbrace{E\left[\left(\overline{(\beta \cdot H_{hl} + \gamma_l)}^k - \overline{(\beta \cdot H_{hl} + \gamma_l)}\right)^2\right]}^{Variance\ between\ tiers}$$
$$+ \underbrace{2Cov((\beta \cdot H_{hl} + \gamma_l) - \overline{(\beta \cdot H_{hl} + \gamma_l)}^k,\ \ \overline{(\beta \cdot H_{hl} + \gamma_l)}^k - \overline{(\beta \cdot H_{hl} + \gamma_l)})}_{Cross\ term}\} + \underbrace{Var(\varepsilon_{hl})}_{variance\ of\ residuals}$$

The sum of the variance within each tier can be decomposed into:

$$\sum_{k=1}^{4} \frac{N^k}{N} \sum_{h=1}^{N^k} \frac{1}{N^k} \{ \overbrace{E\left[\left((\beta \cdot H_{hl} + \gamma_l) - \overline{(\beta \cdot H_{hl} + \gamma_l)}^k\right)^2\right]}^{Sum\ of\ the\ variance\ of\ each\ tier}\}$$

$$= \sum_{k=1}^{4} \frac{N^k}{N} \sum_{h=1}^{N^k} \frac{1}{N^k} \{ \underbrace{E\left[\left((\beta \cdot H_{hl}) - \overline{(\beta \cdot H_{hl})}^k\right)^2\right]}_{household} + \underbrace{E\left[\left(\gamma_l - \overline{\gamma_l}^k\right)^2\right]}_{location} + \underbrace{2Cov\left((\beta \cdot H_{hl}) - \overline{(\beta \cdot H_{hl})}^k, \gamma_l - \overline{\gamma_l}^k\right)}_{interaction} \}$$

The variation between tiers can be decomposed into:

$$\sum_{k=1}^{4} \frac{N^k}{N} \sum_{h=1}^{N^k} \frac{1}{N^k} \{ \overbrace{E\left[\left(\overline{(\beta \cdot H_{hl} + \gamma_l)}^k - \overline{(\beta \cdot H_{hl} + \gamma_l)}\right)^2\right]}^{Variance\ between\ tiers}\}$$

$$= \sum_{k=1}^{4} \frac{N^k}{N} \sum_{h=1}^{N^k} \frac{1}{N^k} \{ \underbrace{E\left[\left(\overline{(\beta \cdot H_{hl})}^k - \overline{(\beta \cdot H_{hl})}\right)^2\right]}_{household} + \underbrace{E\left[\left(\overline{\gamma_l}^k - \overline{\gamma_l}\right)^2\right]}_{location} + \underbrace{2Cov\left(\overline{(\beta \cdot H_{hl})}^k - \overline{(\beta \cdot H_{hl})}, \overline{\gamma_l}^k - \overline{\gamma_l}\right)}_{interaction} \}$$