Industry Agglomeration in a Developing Economy: Evidence from India

Amrit Amirapu[*]
Malavika Thirumalai Ananthakrishnan[†]
Alex Klein[‡]

January 2025

Abstract

This paper examines the strength of agglomeration economies in shaping industrial co-agglomeration in India during the post-liberalisation period. Using granular data from the 2005 Economic Census, we examine the relative importance of two Marshallian channels of agglomeration economies (i.e., input-output linkages and labour market pooling) in determining patterns of industrial concentration, after accounting for natural advantage-driven coagglomeration in India. Employing instrumental variables (based on U.S. industrial data) to address concerns about endogeneity and measurement error, we estimate very large effects for agglomeration economies relative to natural advantages. Our effect sizes are particularly large for both channels relative to what the previous literature has found for advanced economies. Moreover, we find evidence of very distinct dynamics at different geographic scales and for different industrial sectors. We find input-output linkages are more influential at the district level, while labour market pooling is more important at the town/village level. We also find that input-output linkages are far more important for manufacturing industries. Labour market pooling plays a more important role for service sector industries and at finer geographic scales. In contrast, natural advantages appear less influential, though this may reflect measurement challenges. Our study provides a detailed empirical assessment of coagglomeration patterns in a developing country, thus offering insights into the economic forces that shape spatial inequalities.

1 Introduction

Firms benefit from being close to one another. More populated regions offer productivity advantages relative to sparsely populated areas. Moreover, this preference for geographical proximity cannot be explained solely by a region's natural advantages. Scholars agree that a substantial portion of geographical concentration can be attributed to agglomeration externalities, although there is less agreement on which mechanisms are most significant.

We can broadly distinguish between two sets of theories. One argues that natural advantage, often called first nature geography, determines the location of industrial plants. The other

[*] Department of Economics, University of Kent
[†] School of Arts and Science, Azim Premji University
[‡] Department of Economics, University of Sussex

stresses various market mechanisms which incentivise firms to locate near each other (e.g. Ellison and Glaeser 1999). A seminal work by Marshall 1956 argues that labour market pooling and proximity to suppliers and customers can largely explain the formation of geographical clusters. Labour market pooling captures the idea that agglomeration is advantageous for firms because it facilitates better employer-employee matches, while proximity to suppliers and customers captures the idea that locating in geographical clusters allows firms to economise on transport costs when purchasing inputs or delivering goods to downstream customers. Despite a large body of theoretical work examining channels through which labour pooling and proximity of suppliers and customers act as sources of geographical concentration [1], empirical studies have only gradually come to estimate the relative importance of these Marshallian-type externalities. Furthermore, most of the studies have been conducted for advanced economies.

This paper contributes to the growing body of empirical research on agglomeration externalities. Although there is some literature suggesting that agglomeration economies have a productivity-enhancing effect in developing economies (e.g. Lall et al. 2004), little is known about *the sources* of agglomeration economies in developing countries. Given that developing countries differ in terms of their income and productivity levels, consumption patterns, industrial composition, market development, and transport infrastructure, among others, it is not obvious that the drivers of agglomeration will operate as they do in richer countries. Therefore, we ask two questions: Do agglomeration economies matter in developing countries as well? If so, which mechanisms are most important?

We address the questions by investigating the relative importance of labour pooling and proximity to suppliers and customers, respectively, of Indian industries in 2005. For this purpose, we use the methodology developed by Ellison et al. 2010, and combine employment and establishment data at a fine geographic (either district or town/village), and industrial level (4-digit NIC code)[2]. We address concerns regarding reverse causality and measurement error with an instrumental variable strategy based on plausibly exogenous variation in inter-industry linkages and labour market pooling offered by US industry data.

Industrial coagglomeration refers to the extent to which firms from different industries locate near each other. The intuition here is that by locating near firms in different industries, activities can still benefit from proximity through Marshallian interactions, and the reason why they may choose to locate near one another could be due to one or more of the Marshallian channels of agglomeration. By studying coagglomeration patterns, we can better understand the relative importance of the various drivers of agglomeration. For example, firms in different industries may coagglomerate because they share suppliers (input sharing), access the same labour pools, or benefit from overlapping knowledge spillovers. Ellison et al. 2010 developed a methodology to test these economic mechanisms by analysing coagglomeration patterns, which allows for the disentangling of the relative importance of different agglomeration forces. This approach helps explain the drivers of geographical concentration.

Our results can be summarised as follows. First, Marshallian agglomeration economies dominate natural advantages, though the importance of the latter is not trivial. Second, the relative importance of industrial linkages and labour pooling depends on the geographical unit: industrial linkages are more important at the district level, whereas labour pooling at

[1] Lu and Tao 2009; He and Wang 2010; Fernandes and Sharma 2012; Alkay and Hewings 2012; Resende 2015; Mukim 2015; Kathuria 2016; Howard et al. 2016; Barufi et al. 2016; Li and Sinha Roy 2020; Dwivedi and Dubey 2022; Deichmann et al. 2005

[2] The data sets are: the Economic Census, the Annual Survey of Industries, and the Employment-Unemployment Surveys of the National Sample Survey Organization. NIC or National Industry Classification codes are commonly used by statistical agencies in India and roughly correspond to ISIC classification codes.

the city/village level. Third, the strength of agglomeration economies depends on the level of industrial classification in a way that is anticipated: more detailed industrial classification better captures the sources of industrial agglomeration.

The main contribution of the paper is to empirically assess the relative importance of two core agglomeration mechanisms proposed by Marshall 1956, but in the context of a developing economy of India: labour market pooling and proximity to suppliers and customers. This aligns with recent literature that underscores the importance of labour market thickness (Overman and Puga 2010, Miyauchi 2024), and literature on the sources of agglomeration externalities (e.g. Ellison et al. 2010, Kolko 2010, Hanlon and Miscio 2017, Diodato et al. 2018, Stejn et al 2022). However, unlike these studies, our study shifts the emphasis to a developing economy. By doing so, we expand our understanding of how Marshallian agglomeration externalities affect firms' location in different economic conditions than those examined so far. The paper is also related to studies that examine various economic associations with agglomeration patterns in other developing countries (e.g. Deichmann et al. 2005, Alkay and Hewings 2012, Lu and Tao 2009 and He and Wang 2010, Resende 2015, Mukim 2015, Howard et al. 2016).

Naturally, our paper extends literature on spatial development, geographical concentration and agglomeration economies in India (e.g.Amirapu et al. 2019, Duranton, S. E. Ghani, Grover, et al. 2016, Desmet et al. 2013, Rossi-Hansberg 2019). Furthermore, our paper relates to a growing literature on the effects of industrial policies on spatial concentration in developing economies (e.g. E. Ghani et al. 2016, Quin 2017, Li and Sinha Roy 2020, Asher and Novosad 2020, Baum-Snow et al. 2020). Unlike this literature, our paper offers causal mechanisms that explain why firms choose to locate in close proximity to one another, hence helping provide insights into the likely consequences of high local transportation costs.

Our study contributes to several strands in the literature. This is one of the first to empirically assess the determinants of industrial coagglomeration patterns in a developing country at a fine geographic and industrial level.[3] Perhaps the closest study to ours is by Mukim 2015, which examines coagglomeration patterns between formal and informal manufacturing establishments *within* 2-digit industries at the district level in India. Her findings suggest that coagglomeration between formal and informal firms is not very strong, but that the level of coagglomeration that exists is primarily driven by inter-industry input-output and technology linkages. Our study differs in the following ways. First, we aim to explore coagglomeration patterns *between* industries while Mukim 2015 studies coagglomeration between formal and informal establishments *within* industries, so the question is very different. Second, we use data from the 2005 Economic Census (EC) to construct our coagglomeration index rather than the Annual Survey of Industries (ASI). This is important because the ASI is a survey that is only representative at the state level, while the EC is a census and thus representative at any level of geography in theory, allowing us to conduct our analysis at much finer levels of geographic aggregation. Third, our analysis is conducted at the 4-digit industry level, which allows for more precise estimation of inter-industry spillovers than would be possible using coarser measures. Fourth, our data include manufacturing *and service* sector establishments, while most previous studies (including Mukim 2015) focus only on the manufacturing sector. Finally, our identification strategy is different, as we use instrumental variables based on U.S. data, which we argue allows us to estimate the causal effects of our economic mechanisms.

[3] Other studies set in developing countries that also examine the associations between geographical concentration and various economic mechanisms associated with them include: Deichmann et al. 2005 for Indonesia; Alkay and Hewings 2012 for Istanbul; Lu and Tao 2009 and He and Wang 2010 for China; Resende 2015 for Rio de Janerio; Howard et al. 2016 for Vietnam.

The rest of the paper proceeds as follows. Section 2 provides background for the study by briefly describing the economic factors and trends which may have influenced patterns of coagglomeration during the period of study. Section 3 describes the data sources and the construction of the main variables of interest, along with our empirical methodology. Section 4 presents the main results, and Section 5 concludes.

2 Background: Economic Context

This paper aims to shed light on the determinants of coagglomeration patterns in India at a specific moment in time: 2005. In this section, we briefly discuss certain key elements of the Indian economy that are particularly relevant to understanding the context in which this study is situated.

By 2005, the Indian economy was well into its post-liberalisation era, having undergone significant economic reforms starting in the 1980s and 1990s (e.g. de-licensing, trade liberalisation). These reforms dismantled much of the restrictive regulatory framework that had previously governed economic activity, allowing market forces to play a more prominent role. The economy experienced a growth takeoff starting in the early 1980s (Rodrik and Subramanian 2005), and by 2005 the economy was consistently growing at 7% per year. Growth was largely driven by India's services sector, which had become the primary engine of economic expansion (Amirapu and Subramanian 2015; Fan et al. 2023). India's services-led trajectory may have reflected its relatively high levels of investment in tertiary education, which created a skilled workforce capable of supporting knowledge-intensive industries such as information technology and finance (Kochhar et al. 2006). At the same time, the manufacturing sector in India exhibited unusual characteristics for a developing country, being relatively capital-intensive and at the same time constrained by small-scale production, a legacy of post-independence, pre-liberalisation policies that prioritised self-reliance and imposed scale restrictions (Kochhar et al. 2006).

Another notable and relevant feature of the Indian economy post-independence is that it was unusually diversified (Kochhar et al. 2006). Decades of import substitution policies and efforts to achieve industrial self-reliance had fostered a wide variety of industries, even if they did not always exist at globally competitive scales. However, this diversification coexisted with significant structural challenges. Labour mobility, particularly across state boundaries, was limited, partly due to linguistic, cultural, policy and institutional barriers (Kone et al. 2018). This contrasts with other rapidly growing economies like China, where internal migration has been much more pronounced. High congestion costs in urban areas have further compounded the challenges of industrial growth and location choices, as overcrowding in cities has increased the cost of living and doing business.

India's physical infrastructure was another critical factor shaping the economic landscape in 2005. While transportation infrastructure was still underdeveloped compared to global standards, major investments began to take place at the turn of the century to expand rural roads and interstate highways, aiming to improve connectivity within and between urban and rural areas (Asher and Novosad 2020; E. Ghani et al. 2016). Even as progress was underway in reducing the cost of transportation, many frictions have remained. One area in which frictions have been particularly important (especially for the manufacturing sector) is that of land markets, where high transaction costs and unclear property rights (due to inaccurate and outdated land records) have created barriers to efficient land use and acquisition (Duranton, S. E. Ghani, Goswami, et al. 2016; Morris and Pandey 2009). These constraints, combined with high urban congestion costs and limited labour mobility, have helped shape the incentives and disincentives for industries to coagglomerate in India during this period (Amirapu et al. 2019).

3 Data and Methodology

In this study, we use three main data sources: the Fifth Economic Census (EC 2005), the Annual Survey of Industries for 2004/5 (ASI 2005), and Rounds 55, 60 and 61[4] of the Employment and Unemployment Surveys from the National Sample Survey Organization (NSSO).

The EC 2005 is a census of all non-farm establishments[5] in the country. It includes information on establishments' employment levels and industry codes at the 5-digit level of the 2004 National Industry Classification (NIC) system. Being a census, the data are representative of economic activity at the smallest levels of geographic disaggregation by definition. This is essential for estimating coagglomeration at different geographical scales (e.g. village/town, district), and cannot be done with any other available data source. This coverage and representation may come at a cost: measurement error. The EC 2005 is a mammoth data gathering process whose goal is to survey every economic undertaking in the country. In 2005 the Economic Census collected data on almost 42 million establishments (half of which were own-account enterprises). This is quite an achievement for a country with limited financial resources, but it is difficult to achieve with perfect accuracy and relatively untrained enumerators. For this reason, we use an identification strategy that explicitly helps to address this measurement issue. [6]

3.1 Construction of the Coagglomeration Index

We measure industrial coagglomeration following Ellison and Glaeser (1997), henceforth EG, using data from the EC 2005. EG's coagglomeration index measures the extent of geographical concentration between any two industry pairs and aims to capture the strength of agglomerative forces between them. Following EG, we measure agglomeration and coagglomeration using employment as a proxy for economic activity. The coagglomeration index for industry pairs *i* and *j* across *M* regions (indexed by m) is defined as follows:

$$\gamma_{ij}^{c} = \frac{\sum_{m=1}^{M}(s_{mi} - x_m)(s_{mj} - x_m)}{1 - \sum_{m=1}^{M} x_m^2}, \qquad (1)$$

where s_{mi} is the share of industry *i*'s total employment contained in region *m* and x_m is the aggregate size of region m proxied here by region m's share in the country's total employment.[7] From Eq. 1, the coagglomeration index is closely related to the covariance of the region-industry employment shares across industries i and j. The denominator is scaled to eliminate the sensitivity to the geography, controlling the employment distribution across regions. However, it is the numerator that strongly determines the strength of coagglomeration.

[4] These rounds correspond to the years 1999/2000, 2004 and 2004/5, respectively.

[5] *"An establishment would mean a unit/undertaking situated in a single location in which predominantly one economic activity is carried out such that at least some part of the goods/services produced is meant for sale."* (Economic Census, 2013).

[6] The data gathering process was administered at the state level with significant differences in implementation across states. For example, in some states, primary and secondary school teachers were recruited as enumerators, while other states relied on the NSSO's usual stock of trained government enumerators.

[7] When constructing industrial coagglomeration indices among manufacturing industries only, we define x_m as region m's share in total manufacturing employment only.

The value of the EG coagglomeration index ranges from -1 to 1. A higher positive value of the index indicates that the industries are highly geographically concentrated. Any value of the index closer to 0 implies that the industry pairs are spatially dispersed, and negative values of the index imply that the industries are agglomerated in different areas. One limitation of this index is that it measures proximity discretely and does not allow for neighbourhood effects: either two plants are in the same region (e.g. district, town, etc) or they are not.

In our main specification, we measure coagglomeration at the district level using 4-digit (2004 NIC) industry codes. Our data cover 305 4-digit industries and 606 districts. To explore the sensitivity of our results to the scale of geographic and industrial aggregation, we also construct the index at the state as well as the town/village levels and at higher levels of industrial aggregation.[8]

Table 1 lists the 10 most highly coagglomerated industry pairs according to this index constructed at the district level. Some pairs are intuitive (e.g. ”Building and repairing of ships” (3511) and ”Sea and coastal water transport” (6110)) while others are less so (e.g. ”Dressing and dyeing of fur” (1820) and ”Publishing of books” (2211)). The latter may reflect measurement error in the Economic Census as we have described above. Measurement error in an independent variable can lead estimates to be biased *and inconsistent*. To address this concern, we will use an instrumental variable strategy(Cameron and Trivedi 2005, Durbin 1954, etc).

Table 1: Highest Coagglomerated Industries 4-digit: District Level

Industry 1	**Industry 2**	**Coagglomeration**
Manufacture of motorcycles (3591)	Manufacture of bicycles and invalid carriages (3592)	0.1717
Dressing and dyeing of fur; manufacture of articles of fur (1820)	Publishing of books, brochures, musical books, and other publications (2211)	0.1511
Manufacture of Pig Iron (hot metal) (2713)	Manufacture of engines and turbines, except aircraft, vehicle, and cycle engines (2911)	0.1392
Dressing and dyeing of fur; manufacture of articles of fur (1820)	Manufacture of instruments and appliances for measuring, checking, testing, navigating, and other purposes except for industrial process control equipment (3312)	0.1298
Manufacture of machinery for textile, apparel, and leather production (2926)	Manufacture of bicycles and invalid carriages (3592)	0.1206
Scheduled air transport (6210)	Other Business Activities: Market research and public opinion polling (7413)	0.1129
Manufacture of knitted and crocheted fabrics and articles (1730)	Manufacture of bicycles and invalid carriages (3592)	0.1117
Software publishing (7221)	Database activities and distribution of electronic content (7240)	0.1002
Other wholesale (5190)	Sea and coastal water transport (6110)	0.0991
Building and repairing of ships (3511)	Sea and coastal water transport (6110)	0.0990

3.2 Causes of Industrial Coagglomeration

Industries may cluster in a given geographic space for a number of reasons. In this paper, we aim to separately identify the effects of two distinct sources of agglomeration economies: inter-industry input-output linkages and labour market pooling. To do so, we must also separately account for the effect of industrial clustering due to natural advantage, as firms from different industries may choose to co-locate due to nothing more than a shared interest in proximity to a locally available natural advantage. Below, we discuss our construction of proxies for each of these variables as well as our empirical strategy.

[8] At the 2-digit level, we have 59 industries. Tables listing the most highly coagglomerated pairs of industries when using these different levels of geographic and industrial aggregate are provided in Appendix C.3.

It is important to note a caveat here. The ASI data is representative of organised manufacturing industries. Since the input-output linkages are constructed from them, we are able to explore the input linkages only within manufacturing and between manufacturing and services. We are not able to explore input linkages within services.

3.2.1 Intermediary Input Linkage

Inter-industry input-output linkages describe the benefit that firms gain from locating close to potential suppliers of inputs or potential customers of outputs. To construct measures of input-output linkages between industries, we use data on firms' input use from the 2004/5 wave of the Annual Survey of Industries (ASI) to estimate the share of industry i's total inputs that come from industry j: Input$_{i \leftarrow j}$. In most of our analyses, we use a symmetric measure of input-output linkages which captures the maximum input linkage for an industry pair in either direction: Input$_{ij}$ = max(Input$_{i \leftarrow j}$, Input$_{j \leftarrow i}$).

Commodity inputs in the ASI use Annual Survey of Industries Commodity Classification (ASICC) codes, while firm output is classified using NIC codes. We convert the industry-by-commodity pairs into industry-by-industry pairs by writing the 5-digit ASICC codes to 4-digit 2004 NIC codes using an official concordance. The mapping is not 1 to 1, so we omit inputs for which a single ASICC code is mapped to multiple different NIC codes. When multiple ASICC codes are mapped to a single NIC code, we aggregate these ASICC codes together.

3.2.2 Labour Market Pooling

The labour market pooling channel captures the extent to which industries co-locate due to their tendencies to employ similar types of workers. To construct our measures of labour market pooling, we use data from Rounds 55, 60 and 61 of the NSSO's Employment and Unemployment Surveys. These surveys are regionally[9] representative household surveys that include information on individuals' principal economic activities, including the occupations and industries associated with their activities. In all three survey rounds, industries are classified using 2004 NIC industry codes, while occupations are classified using the National Occupation Classification (NOC 1968) system. We use these data to construct an industry (NIC code) by occupation (NOC code) employment matrix, in which each cell of the matrix gives the share of an industry i's total employment held by workers of occupation type o: *share*$_{io}$.[10] We identify individuals engaged in 458 different occupations (identified at NOC 3-digit level). Following Ellison et al. 2010 (henceforth EGK), these data are used to construct a measure of the extent of labour pooling between industry pairs i and j by estimating *corr(share*$_{io}$, share$_{jo}$) across occupations (i.e., the correlation coefficient between *share*$_{io}$ and *share*$_{jo}$). A higher coefficient suggests greater similarity in the occupational structures of industries i and j, and hence higher substitutability between their potential labour pools.

[9] NSS delimited regions do not correspond to standard administrative or political boundaries. They tend to be composed of several districts and are typically smaller than states. There are about 78 ”regions”.

[10] When individuals are employed in multiple economic activities, we use only the industry and occupations associated with the individual's principal activity. The NSS identifies the principal activity of an individual based on the number of hours they are engaged in their activities. If an individual engages in exactly the same hours across different activities, the activity recorded first is considered the principal activity.

3.2.3 Natural Advantage

As discussed above, industries may also co-locate simply due to a shared desire to take advantage of a locally available natural resource (including, for our purposes, labour markets). To account for this source of coagglomeration, we attempt to construct and control for the expected coagglomeration patterns that would occur due to resource and labour market natural advantages alone (denoted by γ_{ij}^{CNA}). We do this, following the method of Ellison and Glaeser 1999, by estimating the predicted spatial distribution of industries across geographical units using only factors associated with natural advantage. Specifically, we regress s_{mi} (the share of industry i's total employment contained in region m) against five industry and region-specific cost advantage factors along with two district-level geographic features river length (RL_m) and coastline (CL_m):

$$s_{mi} = \alpha + \sum_{\zeta} \beta_{\zeta} Y_{\zeta m} Z_{\zeta i} + \beta_{RL} RL_m + \beta_{CL} CL_m + \epsilon_{mi} \tag{2}$$

The cost advantages corresponding to each potential natural resource (ζ) are constructed by taking the product of the average cost of an input in a region ($Y_{\zeta m}$) and the average intensity with which industry uses that input ($Z_{\zeta i}$). The five cost advantage factors are listed in Table 2 below. They cover industry-specific benefits from locating in areas with cheaper electricity, cheaper coal, lower average wages, a larger relative pool of skilled workers, and a larger relative pool of unskilled workers.[11]

Table 2: Data Sources: Effect of Natural Advantages on Region-Industry Employment

Sl. No.	**Region Variable** (Y)	**Source of Region Variable**	**Industry Variable** (Z)	**Source of Industry Variable**
1	Electricity Price	ASI 2005	Electricity Use	ASI 2005
2	Coal Price	ASI 2005	Coal Use	ASI 2005
3	Avg Manuf Wage	ASI 2005	Labour Use	ASI 2005
4	Share Tertiary Edu	NSSO 2004-05	Supervisory and Managerial Staff	ASI 2005
5	Share without degree	NSSO 2004-05	Unskilled and Contract Workers	ASI 2005

After running the regression in equation 2, we use the predicted values ($\hat{s}_{im}$) to construct a coagglomeration index for each industry pair that would be expected due to natural advantages alone:

$$\gamma_{ij}^{CNA} = \frac{\sum_{m=1}^{M} (\hat{s}_{mi} - x_m)(\hat{s}_{mj} - x_m)}{1 - \sum_{m=1}^{M} x_m^2} \tag{3}$$

We then use this term as a control variable in our main analysis to account for observed patterns in coagglomeration that can be explained through natural advantage alone. Ellison and Glaser (1999) in their study show that 20% of the US manufacturing agglomeration can be explained because of local factors. In the study on India, Asher *et al* (2018) find that 10% of the agglomeration is explained because of natural cost differences. One limitation of our study is that we identify only a finite set of natural cost advantages (5 factors) against several others that we are unable to include in the current study. Nevertheless, we show that despite limited factors, our natural advantage variable can explain the spatial concentration.

[11] Details on data construction are provided in Appendix A.1.

3.3 Empirical Strategy: 2SLS Estimation

The main goal of this study is to assess the strength of two potentially important Marshallian channels in predicting the extent of coagglomeration in India. This effort faces two main concerns: endogeneity and measurement error. Measurement error is often a concern when using extremely large datasets from developing countries with lower state capacity and fewer resources for enumerator training and data validation (such as the EC, ASI and NSS datasets used here). Endogeneity may be a concern due to the possibility of reverse causality: Suppose that firms from two industries co-locate in a region due to random chance. These two industries may end up hiring similar workers or using similar inputs merely due to local availability. In that case, there may be an observed relationship between the Marshallian channels and the extent of coagglomeration that is driven by the initial choice to coagglomerate rather than the other way around.

Our empirical strategy is designed to address both concerns by using instrumental variables for our measures of labour pooling and input-output linkages using data from the United States. Because our instruments for these Marshallian channels are generated from non-Indian data, they are less likely to be affected by the same random events that may have led certain industries in India to coagglomerate by chance and are thus less likely to be affected by reverse causality. They may also help with measurement error to the extent that 1) the data quality of the US instruments is higher than the original Indian measures of interest, and 2) the measurement error that exists in the US instruments is uncorrelated with the measurement error in the Indian data.

Our empirical strategy thus consists of using two-stage least squares estimation with the following second-stage regression equation:

$$\gamma_{ij} = \alpha + \beta_{NA}\gamma_{ij}^{CNA} + \beta_L Labo\hat{u}rCor_{ij} + \beta_{IO} Inp\hat{u}t_{ij} + \epsilon_{ij} \tag{4}$$

Here, γ_{ij} represents the pairwise coagglomeration index value for industries i and j; γ_{ij}^{CNA} is the pairwise coagglomeration index for industries i and j predicted by the industries' exposure to natural advantages alone; $Inp\hat{u}t_{ij}$ and $Labo\hat{u}rCor_{ij}$ are the predicted values of the two main Marshallian channels (input-output linkages and labour market pooling) using the US instruments discussed above.

3.3.1 US Instrument Construction

Data on intermediary input linkages for US industries (i.e. $Input_{ij}^{US}$.) are obtained from US Input-Output Tables for 2007.[12] The relevant table identifies 405 industries using the 2007 NAIC (North American Industry Classification System) system at the 6-digit level. Since there is no direct concordance between (Indian) NIC and (US) NAIC codes, we made a mapping using a series of intermediate concordances: we first mapped the 2007 NAIC codes to 2002 NAIC codes, which were then mapped to ISIC (International Standard Industrial Classification) 3.1 codes, which were finally mapped to 4-digit NIC codes. This left us with only 123 unique NIC codes at the 4-digit level for which there is data on input-output linkages from both the US and India.

The data used to construct our instrument for labour pooling are taken from the 2014 US National Industry-Occupation Employment Matrix (NIOEM). These data are available for 455 industries at the 6-digit 2012 NAIC level across 1090 occupations. To map the 2012 NAIC

[12] We construct the input coefficient matrix following the commodity-by-industry Use Table (2007), which gives the uses of commodities by intermediate and final goods users.

codes to 4-digit 2004 NIC codes, we first mapped the 2012 NAIC codes to 2007 NAIC codes and then used the mapping discussed directly above. This process left us with 183 unique 4-digit 2004 NIC codes for which there were data from both the US and India. We then performed a similar exercise to that described in Section 3.2.2 to generate a measure of the extent to which two industries share similar occupational structures in the US: $LabourCorr_{ij}^{US}$.

Table 3: Summary Statistics 4-Digit District level

	Full Sample			
	Mean	Standard deviation	Minimum	Maximum
EG Coagg. Index	0.0010	0.0060	-0.0102	0.1717
Nat.-Adv. EG Coagg. Index	0.0003	0.0002	-0.0005	0.0019
Input-Output	0.0087	0.0517	0.0000	1.0000
Labour Correlation	0.1503	0.5221	-1.0000	1.0000
Observations	9854			

Table 4: Summary Statistics 4-Digit District level

	IV Sample			
	Mean	Standard deviation	Minimum	Maximum
EG Coagg. Index	0.0013	0.0044	-0.0077	0.0449
Nat.-Adv. EG Coagg. Index	0.0004	0.0002	-0.0002	0.0012
Input-Output	0.0118	0.0578	0.0000	0.9914
Labour Correlation	0.1891	0.4347	-1.0000	1.0000
Input-Output (US)	0.0146	0.0430	0.0000	0.6785
Labour Correlation (US)	0.8436	0.0906	0.5065	0.9952
Observations	2258			

4 Empirical Results

In this section, we present the main results from estimating equation 4. Again, the goal is to estimate the effects of our two Marshallian channels of interest on the extent of industrial coagglomeration in India. In our main specifications, we measure coagglomeration between 4-digit industries at the district level. In the Appendix, we present results at other levels of aggregation (i.e. 2-digit industry). Our coagglomeration index is constructed using data on employment from a census of enterprises that includes all formal *and informal* establishments in both manufacturing *and services* industries.

The main results are presented in Table 5, where all the variables are standardised to have zero mean and unit standard deviation to facilitate comparison of effect sizes across variables and studies. In the first column, we present results using OLS. The coefficients on our measures of input-output linkages and natural advantages are positive but modest, while they are slightly negative for the case of labour pooling. Our primary specification - shown in column 3 -

includes instrumental variables from the US to address the potential problems of endogeneity and measurement error discussed above.

The results from our main specification (column 3 of Table 5) suggest large roles for both labour pooling and input-output linkages in determining coagglomeration patterns: a one standard deviation increase in our measure of input-output linkages (labour pooling) for an industry pair is associated with an increase in observed coagglomeration (at the 4 digit industry x district level) of .30 (.44)[13] standard deviations. These coefficients are particularly large in comparison with those from previous studies, such as EGK: in the primary specification of their study of coagglomeration patterns in the US (column 1 of Table 4), they find effect sizes of .15 and .12 for input linkages and labour pooling, respectively.

By contrast, the coefficient on our measure of natural advantages is small (.03) and insignificant - especially in comparison to the large effect found by EGK (.16) - though this could be due to limitations in our ability to measure natural advantages with high accuracy and precision. Nonetheless, we find that intermediary input linkages are strong drivers of industrial coagglomeration at the district level, and the effect is stronger at granular industrial levels.

One immediate concern when comparing the OLS results with the IV results is the considerable reduction in sample size: from 9,854 observations in column 1, we retain only 2,258 in column 3. This is due to the fact that there are many industry pairs for which we do not have data on US instruments (see Section 3.3.1 above). To see whether the differential effects observed in column 3 are due to the instruments or the different sample, we reproduce the OLS results on the restricted IV sample in Column 2. Reassuringly, we find that the effects are very similar between Columns 1 and 2, suggesting that the differences we observe in Column 3 are *not* due merely to a difference in the sample of available industries and are, therefore, likely to be caused by a reduction in the bias generated by measurement error and reverse causality [14]

In Table 6 we present the results from our first-stage regressions. The results show that our two instruments for labour pooling and input-output linkages are strong, with F-statistics greater than 10 for both instruments. But we also report the weak-instrument robust Andreson-Rubin CI (confidence interval) for the endogenous variables. We argued that the IVs satisfy the relevance and exclusion criteria and that the IVs are also unlikely to reflect reverse causality between observed coagglomeration patterns in India and Marshallian linkages because they reflect industry relationships from another country and should therefore be unrelated to any past events that caused certain industries to coagglomerate through random chance in India.

In addition to estimating specification 4 using data constructed at the 4-digit NIC x district level, we also do so at other levels of aggregation. Table 7 presents results when constructing the coagglomeration index at an even finer geographic level: the town/village level.[15] The results suggest a similar role for input-output linkages, for which the coefficient is comparable with the observed effect at the district level. However, the effect of labour pooling seems to

[13] In this specification the coefficient on the labour pooling channel just misses significance at the 10%level, but the results are nonetheless suggestive of a potentially large effect.

[14] The coefficients and standard errors from the IV estimates are larger than those from the corresponding OLS estimates for input-output linkages. To show that IV corrects for the bias in the OLS model, we conduct a paired t-test comparing the coefficients from the two models. The results, presented in Tables (16 and 17), show that the coefficients from the IV and OLS models are statistically significantly different. This test is performed for both OLS models (for the full sample as well as the IV sample) in comparison to the IV estimates, and the results consistently confirm the significant difference.

[15] Results from estimating specification 4 at coarser levels of industrial aggregation (i.e. 2-digit NIC code) are provided in the appendix. Briefly, they suggest that input-output linkages are the strongest determinants of coagglomeration patterns at the 2-digit level, while labour pooling and natural advantages are not significant factors.

Table 5: Main results (4-digit NIC code and district level)

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.068***	0.077***	0.295***
	(0.018)	(0.026)	(0.098)
Anderson-Rubin 90% confidence interval			[.135353, .645675]
Labour Correlation	-0.031**	-0.041**	0.441
	(0.014)	(0.021)	(0.270)
Anderson-Rubin 90% confidence interval			[-.088616, 1.0582]
Natural Advantages	0.021*	0.042***	0.026
	(0.011)	(0.015)	(0.018)
Sample	Full	IV	IV
Kleibergen–Paap F stat			18.53
Anderson-Rubin Chi-sq			16.29
Observations	9854.000	2258.000	2258.000

Note: This table presents the results estimating the effects of Marshallian channels of intermediary input-output linkages and labour market pooling on industrial coagglomeration of 4-digit industries at the district level in India for 2005. The OLS results present bootstrapped standard errors in parentheses clustered at the three-digit industry pairs, whereas the IV results present clustered standard errors. Stars indicate statistical significance: **p <0.10, ** p <0.05, *** p <0.01*. All variables are standardised to have a mean of zero and unit standard deviation. Column 1 presents OLS results for the full sample. Column 2 presents OLS results for the IV sample. Column 3 presents IV results. *[AR sig = 0.0003]*

Table 6: First-Stage of IV (4-digit NIC code)

	(1) Input-Output	(2) Labour Correlation
Labour Correlation (US)	0.044	0.141***
	(0.040)	(0.022)
Input-Output (US)	0.635***	-0.034*
	(0.108)	(0.014)
Natural Advantages	0.030	-0.028
	(0.021)	(0.015)
F-Statistic	18.63	19.08
Observations	2258	2258

Note: This table presents the first-stage results following the IV strategy for all 4-digit industries. Each column presents the separate first-stage results for the two endogenous variables capturing the Marshallian channels of input linkages and labour pooling instrumented using US data. Standard errors in parentheses. Stars indicate statistical significance: **p <0.10, ** p <0.05, *** p <0.01*. All variables are standardised to have a mean of zero and a unit standard deviation.

be much greater at this finer level: a one standard deviation increase in our measure of labour pooling is associated with an increase in observed coagglomeration (at the 4 digit industry x town level) of 1.00 standard deviation - suggesting that labour markets are very important in determining industrial location decisions but are also geographically constricted to nearby areas. This implies that the costs of transporting workers across moderately large distances (i.e. across towns within districts) in India are particularly high. Now, it should be emphasised that the data, when constructed at the town/village level, may systematically exclude large cities[16] - in which the costs of commuting may be substantially lower, so these results should be understood as speaking to transport costs outside of large metropolitan areas.

Although the data that we use to generate the coagglomeration indices include all non-

[16] This is a necessary consequence of how the SHRUG dataset was put together. See Asher et al. 2021 for further information.

Table 7: Main results (4-digit NIC code and town level)

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.048**	0.049*	0.217***
	(0.019)	(0.029)	(0.063)
Anderson-Rubin 90% confidence interval			[.134352, .464457]
Labour Correlation	-0.015	-0.007	0.999***
	(0.017)	(0.029)	(0.304)
Anderson-Rubin 90% confidence interval			[.601121, 1.79431]
Natural Advantages	0.041***	0.053***	0.040*
	(0.012)	(0.017)	(0.023)
Sample	Full	IV	IV
Kleibergen–Paap F stat			18.53
Anderson-Rubin Chi-sq			29.52
Observations	9854.000	2258.000	2258.000

Note: This table presents the results estimating the effects of Marshallian channels of intermediary input-output linkages and labour market pooling on industrial coagglomeration of 4-digit industries at the village/town/city level in India for 2005. The OLS results present bootstrapped standard errors in parentheses clustered at the three-digit industry pairs, whereas the IV results present clustered standard errors. Stars indicate statistical significance: **p <0.10, ** p <0.05, *** p <0.01*. All variables are standardised to have a mean of zero and unit standard deviation. Column 1 presents OLS results for the full sample. Column 2 presents OLS results for the IV sample. Column 3 presents IV results. *[AR p val = 0.0001]*

farm enterprises and comprise manufacturing and services, the same cannot be said of some of our independent variables. Specifically, our measure of input-output linkages is generated from the ASI, a representative survey of the formal manufacturing sector, and thus does not (generally) include service sector establishments. Our measure of natural advantages also relies primarily on data from the ASI and is thus unlikely to be able to explain coagglomeration patterns between industries in the service sector. For this reason, we run a version of our main specification (equation 4) at the 4-digit industry x district level using only data from the manufacturing sector. The results are presented in Table 8 below. Focusing on the IV results in column 3, we see that the effect of input-output linkages is similar to what was found in Table 5. The coefficient on the labour pooling measure, however, is no longer statistically significant - and, in fact, the first stage of the IV estimation for labour pooling (Table 12 in the Appendix), suggests that this instrument is weak when using only data from the manufacturing sector. To address this, we report weak instrument robust Anderson-Rubin confidence intervals for the endogenous variables. What we find is that much of the relevant variation in labour pooling comes from establishments in the service sector rather than manufacturing. The natural advantage term remains small and statistically insignificant.

5 Conclusion

This study aims to provide the most comprehensive examination yet of the sources driving geographical concentration in the Indian non-farm sector post-liberalisation. By leveraging the methodologies developed by Ellison, Glaeser, and others, and by utilising a variety of different large-scale data sources, we uncover a significant role for inter-industry linkages and labour market pooling across manufacturing and services as critical factors in explaining industrial coagglomeration in India. Our instrumental variable strategy, which uses exogenous variation from US industry data, mitigates concerns of reverse causality and measurement error, thereby

Table 8: Manufacturing Sector (4-digit NIC code and district level)

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.065***	0.085***	0.215***
	(0.017)	(0.025)	(0.081)
Anderson-Rubin 90% confidence interval			[.057115, .478214]
Labour Correlation	-0.044***	-0.053**	-0.159
	(0.016)	(0.026)	(0.670)
Anderson-Rubin 90% confidence interval			[-1.90864, 1.59152]
Natural Advantages	-0.012	0.027	0.011
	(0.029)	(0.046)	(0.062)
Sample	Full	IV	IV
Kleibergen–Paap F stat			4.66
Anderson-Rubin Chi-sq			7.40
Observations	7099.000	1465.000	1465.000

Note: This table presents the results estimating the effects of Marshallian channels of intermediary input-output linkages and labour market pooling on industrial coagglomeration of 4-digit industries at the district level in India for 2005, for only Manufacturing industries. The OLS results present bootstrapped standard errors in parentheses clustered at the three-digit industry pairs, whereas the IV results present clustered standard errors. Stars indicate statistical significance: **p <0.10, ** p <0.05, *** p <0.01*. All variables are standardised to have a mean of zero and unit standard deviation. Column 1 presents OLS results for the full sample. Column 2 presents OLS results for the IV sample. Column 3 presents IV results. *[AR P-val= 0.02.]*

strengthening the validity of our findings.

The results reveal that while inter-industry input-output linkages consistently play a substantial role in determining coagglomeration patterns at most levels of aggregation, labour market pooling is especially important in explaining coagglomeration at the smallest spatial units (i.e. town/village level) and seems most relevant for service sector establishments only. Natural advantages, by contrast, are not found to play a significant role in this context, although this may reflect difficulties in measuring them precisely. This nuanced understanding challenges some conventional views and highlights the unique dynamics within developing economies like India. The different results for labour pooling across different spatial distances suggest that high transport costs may lower aggregate productivity by increasing the cost of labour migration and commuting potential. Our analysis also suggests that policies which aim at enhancing industrial linkages may be more effective in promoting productive agglomeration decisions when that is the goal of the states. We find that at granular scales, the agglomeration channels are much stronger.

Overall, our study adds to the literature on agglomeration economies by providing new insights into the mechanisms underlying industrial concentration in a developing country. It also offers valuable implications for policymakers aiming to foster economic growth and reduce spatial inequalities through targeted infrastructure and industrial policies. The findings suggest that facilitating stronger inter-industry connections and reducing the cost of commuting and migration within districts could help leverage the productivity benefits of agglomeration in developing country contexts.

References

Alkay, Elif and Geoffrey JD Hewings (2012). “The determinants of agglomeration for the manufacturing sector in the Istanbul metropolitan area”. In: *The Annals of regional science* 48, pp. 225–245.

Amirapu, Amrit, Rana Hasan, Yi Jiang, and Alex Klein (2019). “Geographic concentration in Indian manufacturing and service industries: Evidence from 1998 to 2013”. In: *Asian Economic Policy Review* 14.1, pp. 148–168.

Amirapu, Amrit and Arvind Subramanian (2015). *Manufacturing or Services? An Indian Illustration of a Development Dilemma*. Working Paper 408. Center for Global Development. DOI: 10.2139/ssrn.2623158. URL: https://ssrn.com/abstract=2623158.

Asher, Sam, Tobias Lunt, Ryu Matsuura, and Paul Novosad (2021). “Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the shrug open data platform”. In: *The World Bank Economic Review* 35.4, pp. 845–871.

Asher, Sam and Paul Novosad (2020). “Rural roads and local economic development”. In: *American economic review* 110.3, pp. 797–823.

Barufi, Ana Maria Bonomi, Eduardo Amaral Haddad, and Peter Nijkamp (2016). “Industrial scope of agglomeration economies in Brazil”. In: *The Annals of Regional Science* 56, pp. 707–755.

Baum-Snow, Nathaniel, J Vernon Henderson, Matthew A Turner, Qinghua Zhang, and Loren Brandt (2020). “Does investment in national highways help or hurt hinterland city growth?” In: *Journal of Urban Economics* 115, p. 103124.

Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Deichmann, Uwe, Kai Kaiser, Somik V Lall, and Zmarak Shalizi (2005). “Agglomeration, transport, and regional development in Indonesia”. In: *Available at SSRN 647682*.

Desmet, Klaus, Ejaz Ghani, and Stephen O’Connell (2013). “India’s Spatial Development”. In: *World Bank-Economic Premise* 124, pp. 1–5.

Diodato, Dario, Frank Neffke, and Neave O’Clery (2018). “Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time”. In: *Journal of Urban Economics* 106, pp. 1–26.

Duranton, Gilles, Syed Ejaz Ghani, Arti Grover Goswami, and William Robert Kerr (Jan. 2016). *A detailed anatomy of factor misallocation in India*. Policy Research Working Paper Series 7547. The World Bank. URL: https://ideas.repec.org/p/wbk/wbrwps/7547.html.

Duranton, Gilles, Syed Ejaz Ghani, Arti Grover, William Kerr, and William Robert Kerr (2016). “A detailed anatomy of factor misallocation in India”. In: *World Bank Policy Research Working Paper* 7547.

Durbin, J. (1954). “Errors in Variables”. In: *Revue de l’Institut International de Statistique / Review of the International Statistical Institute* 22.1/3, pp. 23–32. ISSN: 03731138. URL: http://www.jstor.org/stable/1401917 (visited on 01/06/2025).

Dwivedi, Aasheerwad and Amaresh Dubey (2022). “Coagglomeration And New Economic Activity: Evidence From Formal And Informal Manufacturing Firms”. In: *Journal of Developmental Entrepreneurship* 27.02, p. 2250013.

Ellison, Glenn and Edward L Glaeser (1999). “The geographic concentration of industry: does natural advantage explain agglomeration?” In: *American Economic Review* 89.2, pp. 311–316.

Ellison, Glenn, Edward L Glaeser, and William R Kerr (2010). "What causes industry agglomeration? Evidence from coagglomeration patterns". In: *American Economic Review* 100.3, pp. 1195–1213.

Fan, Tianyu, Michael Peters, and Fabrizio Zilibotti (2023). "Growing Like India—the Unequal Effects of Service-Led Growth". In: *Econometrica* 91.4, pp. 1457–1494. DOI: https://doi.org/10.3982/ECTA20964. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA20964. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA20964.

Fernandes, A and Gunjan Sharma (2012). *The role of infrastructure, governance, education and industrial policy*. Tech. rep. Working Paper No. C-35005-INC-1). International Growth Centre.

Ghani, Ejaz, Arti Grover Goswami, and William R. Kerr (Apr. 2016). "Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing". In: *The Economic Journal* 126.591, pp. 317–357. ISSN: 0013-0133. DOI: 10.1111/ecoj.12207. eprint: https://academic.oup.com/ej/article-pdf/126/591/317/25830031/ecoj12207-sup-0001-appendixa.pdf. URL: https://doi.org/10.1111/ecoj.12207.

Hanlon, W Walker and Antonio Miscio (2017). "Agglomeration: A long-run panel data approach". In: *Journal of Urban Economics* 99, pp. 1–14.

He, Canfei and Junsong Wang (2010). "Geographical agglomeration and co-agglomeration of foreign and domestic enterprises: A case study of Chinese manufacturing industries". In: *Post-Communist Economies* 22.3, pp. 323–343.

Howard, Emma, Carol Newman, and Finn Tarp (2016). "Measuring industry coagglomeration and identifying the driving forces". In: *Journal of Economic Geography* 16.5, pp. 1055–1078.

Kathuria, Vinish (2016). "What Causes Agglomeration—Policy or Infrastructure? A Study of Indian Organised Manufacturing". In: *Economic and Political Weekly*, pp. 33–44.

Kochhar, Kalpana, Utsav Kumar, Raghuram Rajan, Arvind Subramanian, and Ioannis Tokatlidis (July 2006). "India's pattern of development: What happened, what follows?" In: *Journal of Monetary Economics* 53.5, pp. 981–1019. URL: https://ideas.repec.org/a/eee/moneco/v53y2006i5p981-1019.html.

Kolko, Jed (2010). "Urbanization, agglomeration, and coagglomeration of service industries". In: *Agglomeration economics*. University of Chicago Press, pp. 151–180.

Kone, Zovanga L, Maggie Y Liu, Aaditya Mattoo, Caglar Ozden, and Siddharth Sharma (2018). "Internal borders and migration in India". In: *Journal of Economic Geography* 18.4, pp. 729–759. URL: https://ideas.repec.org/a/oup/jecgeo/v18y2018i4p729-759..html.

Lall, Somik V, Zmarak Shalizi, and Uwe Deichmann (2004). "Agglomeration economies and productivity in Indian industry". In: *Journal of development economics* 73.2, pp. 643–673.

Li, Yue and Sutirtha Sinha Roy (2020). "The Employment Effect of Place-Based Policies". In.

Lu, Jiangyong and Zhigang Tao (2009). "Trends and determinants of China's industrial agglomeration". In: *Journal of urban economics* 65.2, pp. 167–180.

Marshall, Alfred (1956). *Principles of Economics 8th Edition,(1920), 10th reprinted*.

MIT (n.d.). *MIT Geodata Repository*. URL: https://geodata.mit.edu.

Miyauchi, Yuhei (2024). "Matching and Agglomeration: Theory and Evidence From Japanese Firm-to-Firm Trade". In: *Econometrica* 92.6, pp. 1869–1905.

Morris, Sebastian and Ajay Pandey (2009). "Land Markets in India: Distortions and Issues". In: *India Infrastructure Report 2009: Land - A Critical Resource for Infrastructure*. Ed. by

Nirmal Mohanty, Runa Sarkar, and Ajay Pandey. Available at SSRN: https://ssrn.com/abstract=2182490. New Delhi: Oxford University Press, chapter pagination if available.

Mukim, Megha (2015). "Coagglomeration of formal and informal industry: evidence from India". In: *Journal of Economic Geography* 15.2, pp. 329–351.

Overman, Henry G and Diego Puga (2010). "Labor pooling as a source of agglomeration: An empirical investigation". In: *Agglomeration economics*. University of Chicago Press, pp. 133–150.

Resende, Marcelo (2015). "Industrial coagglomeration: some state-level evidence for Brazil". In: *Nova Economia* 25, pp. 181–194.

Rodrik, Dani and Arvind Subramanian (2005). "From "Hindu Growth" to Productivity Surge: The Mystery of the Indian Growth Transition". In: *IMF Staff Papers* 52.2, pp. 193–228.

Rossi-Hansberg, Esteban (2019). "Geography of growth and development". In: *Oxford Research Encyclopedia of Economics and Finance*.

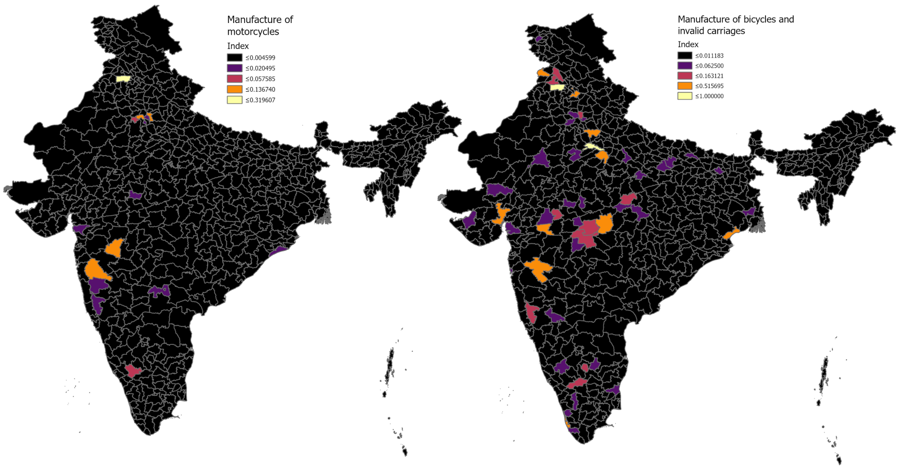
Appendix

A.1 Data Background

A.1.1 Natural Advantages

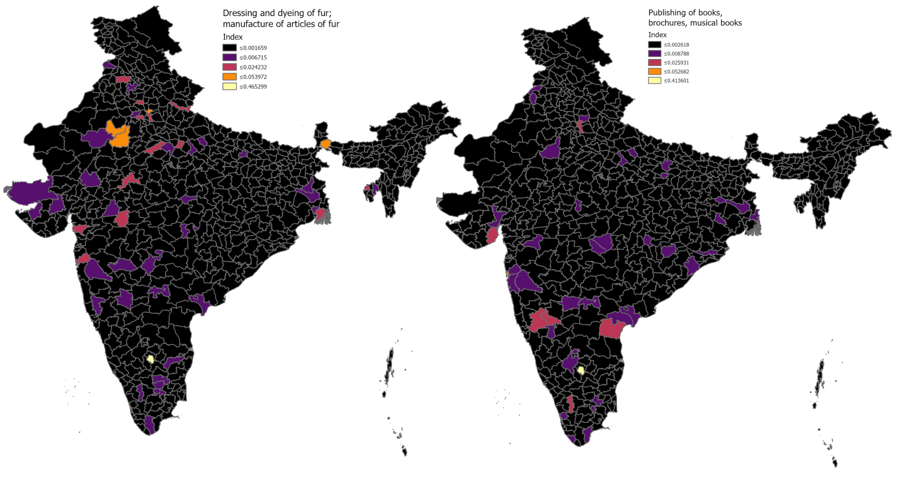
Prices of electricity, coal and average manufacturing wages are computed from firm-level ASI data averaged across firms in a district. Industry-level usage of these inputs and factors (i.e. electricity, coal and labour) is calculated by taking the ratio of total cost from that input or factor over total gross value added, averaged across industries. The availability of skilled labour is proxied by the share of the population in a district with tertiary degrees and above and is generated from NSSO data. The share of unskilled workers - also calculated from NSSO data - is the share of the total population with educational attainment below high school completion. The corresponding industry intensities for these two variables are calculated using ASI data by taking the fraction of workers in an industry who belong to supervisory and managerial staff (to proxy for human capital intensity) and the fraction of workers in an industry who are labourers hired directly by firms or through contractors (to proxy for unskilled labour intensity). River length (RL_m) and coastline (CL_m) are computed from ArcGIS Pro using shape files for the year 2002 from the MIT Geo repository MIT n.d.

B.2 Additional Figures



The map presents the shares of activities of two industries: The manufacture of motorcycles (3591) and the Manufacture of bicycles and invalid carriages (3592), which represent the highest coagglomerated industries in India at the disaggregated industrial classification of 0.1716.

Figure 1: Geographical Concentration of Economic Activity



The map presents the shares of activities of two industries: Dressing and dyeing of fur; manufacture of articles of fur (1820) and Publishing of books, brochures, musical books and other publications (2211), which represents the highest coagglomerated industries in India at the disaggregated industrial classification of 0.1511.

Figure 2: Geographical Concentration of Economic Activity

C.3 Additional Tables

Table 9: Highest Coagglomerated Industries 4-digit: State Level

Industry 1	Industry 2	Coagglomeration
Manufacture of knitted and crocheted fabrics and articles 1730	Manufacture of other chemical product n.e.c 2429	0.3706364
Hunting, trapping and game propagation, including related service activities 150	Manufacture of Pig Iron 2713	0.320511
Manufacture of aircraft and spacecraft 3530	Scheduled air transport 6210	0.2910531
Manufacturing of wooden containers 2023	Manufacture of semi-finished products (2714	0.2648426
Manufacture of semi-finished products 2714	Other recreational activities 9249	0.2542925
Manufacture of aircraft and spacecraft 3530	Botanical and zoological gardens and nature reserves activities. 9233	0.2458238
Mining of iron ores 1310	Manufacture of semi-finished products 2714	0.2433833
Manufacture of aircraft and spacecraft 3530	Other credit granting 6592	0.2432815
Hunting, trapping and game propagation, including related service activities 150	Sea and coastal water transport 6110	0.2383357
Mining and agglomeration of lignite 1020	Wholesale of metals and metal ores 5142	0.2319463

Table 10: Highest Coagglomerated Industries 4-digit: Town/Village Level

Industry 1	Industry 2	Coagglomeration
Manufacture of instruments and appliances for measuring, checking, testing, navigating and other purposes except for industrial process control equipment 3312	Publishing of books, brochures, musical books and other publications 2211	0.6307989
Dressing and dyeing of fur; manufacture of articles of fur 1820	Manufacture of instruments and appliances for measuring, checking, testing, navigating and other purposes except for industrial process control equipment 3312	0.4515104
Manufacture of other articles of paper and paperboard 2109	Architectural and engineering activities and related technical consultancy[7421	0.3678543
Manufacture of other articles of paper and paperboard 2109	Wholesale of textiles, clothing and footwear 5131	0.3312897
Manufacture of insulated wire and cable 3130	Wholesale of textiles, clothing and footwear 5131	0.3146717
Dressing and dyeing of fur; manufacture of articles of fur 1820	Manufacture of insulated wire and cable 3130	0.3089106
Publishing of books, brochures, musical books and other publications 2211	Manufacture of domestic appliances, n.e.c 2930	0.2981295
Dressing and dyeing of fur; manufacture of articles of fur 1820	Architectural and engineering activities and related technical consultancy 7421	0.2828872
Dressing and dyeing of fur; manufacture of articles of fur 1820	Real estate activities on a fee or contract basis 7020	0.281903

Table 11: Highest Coagglomerated Industries 2-digit: District Level

Industry 1	Industry 2	Coagglomeration
Manufacture Of Office, Accounting And Computing Machinery-30	Manufacture Of Radio, Television And Communication Equipment And Apparatus-32	0.0338063
Manufacture Of Office, Accounting And Computing Machinery-30	Manufacture Of Motor Vehicles, Trailers And Semi-Trailers-34	0.0313395
Manufacture Of Office, Accounting And Computing Machinery-30	Research And Development-73	0.0307583
Extraction Of Crude Petroleum And Natural Gas; Service Activities Incidental To Oil And Gas Extraction Excluding Surveying-11	Air Transport-62	0.0285639
Tanning And Dressing Of Leather; Manufacture Of Luggage, Handbags Saddlery, Harness And Footwear-19	Water Transport-61	0.0256339
Manufacture Of Office, Accounting And Computing Machinery-30	Computer And Related Activities-72	0.0244869
Mining Of Uranium And Thorium Ores-12	Manufacture Of Tobacco Products-16	0.01926
Water Transport-61	Air Transport-62	0.018204
Manufacture Of Medical, Precision And Optical Instruments, Watches And Clocks-33	Computer And Related Activities-72	0.017988
Extraction Of Crude Petroleum And Natural Gas; Service Activities Incidental To Oil And Gas Extraction Excluding Surveying-11	Water Transport-61	0.0155427

Table 12: Manufacturing Sector: First-Stage of IV (4-digit NIC code)

	(1) Input-Output	(2) Labour Correlation
Labour Correlation (US)	0.051	0.112**
	(0.077)	(0.038)
Input-Output (US)	0.703***	-0.029
	(0.117)	(0.015)
Natural Advantages	0.106*	0.079*
	(0.046)	(0.037)
F-Statistic	19.58	5.29
Observations	1465	1465

Note: This table presents the first-stage results following the IV strategy for 4-digit manufacturing industries. Each column presents the separate first-stage results for the two endogenous variables capturing the Marshallian channels of input linkages and labour pooling instrumented using US data. Standard errors in parentheses are clustered at the three-digit industry level. Stars indicate statistical significance: **p <0.10, ** p <0.05, *** p <0.01.* All variables are standardised to have a mean of zero and a unit standard deviation.

Table 13: First-Stage of IV (2-digit NIC code)

	(1) Input-Output	(2) Labour Correlation
Labour Correlation (US)	0.013	0.352***
	(0.031)	(0.083)
Input-Output (US)	0.327*	-0.093*
	(0.151)	(0.043)
Natural Advantages	0.039	0.011
	(0.054)	(0.063)
F-Statistic	2.95	11.32
Observations	428	428

Note: This table presents the first-stage results following the IV strategy for 2-digit industries. Each column presents the separate first-stage results for the two endogenous variables capturing the Marshallian channels of input linkages and labour pooling instrumented using US data. Standard errors in parentheses are clustered at the one-digit industry level. Stars indicate statistical significance: **p <0.10, ** p <0.05, *** p <0.01.* All variables are standardised to have a mean of zero and a unit standard deviation.

Table 14: 2-digit NIC code and district level

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.125	0.215	1.278
	(0.115)	(0.214)	(0.828)
Labour Correlation	-0.045	-0.135	0.084
	(0.068)	(0.109)	(0.272)
Natural Advantages	-0.012	0.066	-0.042
	(0.072)	(0.104)	(0.089)
Sample	Full	IV	IV
R^2	0.012	0.035	-0.522
Kleibergen–Paap F stat			3.62
Anderson-Rubin Chi-sq			4.73
Observations	649.000	428.000	428.000

Note: This table presents the results estimating the effects of Marshallian channels of intermediary input-output linkages and labour market pooling on industrial coagglomeration of 2-digit industries at the district level in India for 2005. The OLS results present bootstrapped standard errors in parentheses clustered at the one-digit industry pairs, whereas the IV results present clustered standard errors. Stars indicate statistical significance: **p <0.10, ** p <0.05, *** p <0.01.* All variables are standardised to have a mean of zero and unit standard deviation. Column 1 presents OLS results for the full sample. Column 2 presents OLS results for the IV sample. Column 3 presents IV results. AR significance - 0.09

Table 15: 2-digit NIC code and Town/Village level

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.059	0.146	0.928
	(0.079)	(0.136)	(0.621)
Labour Correlation	-0.046	-0.148	0.012
	(0.079)	(0.128)	(0.287)
Natural Advantages	0.076	0.128	0.048
	(0.061)	(0.086)	(0.083)
Sample	Full	IV	IV
R^2	0.009	0.038	-0.307
Kleibergen–Paap F stat			3.62
Anderson-Rubin Chi-sq			5.39
Observations	649.000	428.000	428.000

Note: This table presents the results estimating the effects of Marshallian channels of intermediary input-output linkages and labour market pooling on industrial coagglomeration of 2-digit industries at the village/town/city level in India for 2005. The OLS results present bootstrapped standard errors in parentheses clustered at the one-digit industry pairs, whereas the IV results present clustered standard errors. Stars indicate statistical significance: **p <0.10, ** p <0.05, *** p <0.01*. All variables are standardised to have a mean of zero and unit standard deviation. Column 1 presents OLS results for the full sample. Column 2 presents OLS results for the IV sample. Column 3 presents IV results. AR - p val = 0.06

Table 16: Two-sample t-test: OLS and IV results

	Observations	Mean	Std. err.	Std. dev.	[95% conf. interval]
OLS	9,854	0.0680421	0.0001501	0.0149042	[0.0677478, 0.0683364]
IV	2,258	0.2948285	0.0020114	0.0955769	[0.2908842, 0.2987728]
Combined	12,112	0.1103211	0.0008942	0.0984101	[0.1085684, 0.1120739]
Difference		-0.2267864	0.0010125		[-0.2287711, -0.2248017]
t-statistic					-2.2e+02
Degrees of freedom					12,110
P-values	Ha: diff < 0	Ha: diff != 0	Ha: diff > 0		
	0.0000	0.0000	1.0000		

Note: The table presents the results of the T-test examining if the coefficients of the Input-output linkages following 5 of OLS and the IV strategy are statistically significantly different from one another. The results compare the coefficients and standard errors of columns (1) and (3). The results of the paired t-test suggest that the observed coefficients on input linkages from IV and OLS are statistically significantly different.

Table 17: Two-sample t-test: OLS of IV sample and IV results

	Observations	Mean	Std. err.	Std. dev.	[95% conf. interval]
OLS of IV sample	2,258	0.0772254	0.0005186	0.0246423	[0.0762084, 0.0782424]
IV	2,258	0.2948285	0.0020114	0.0955769	[0.2908842, 0.2987728]
Combined	4,516	0.1860269	0.0019236	0.1292688	[0.1822557, 0.1897982]
Difference		-0.2176031	0.0020771		[-0.2216753, -0.2135309]
t-statistic					-1.0e+02
Degrees of freedom					4,514
P-values	Ha: diff < 0	Ha: diff != 0	Ha: diff > 0		
	0.0000	0.0000	1.0000		

Note: The table presents the results of the T-test examining if the coefficients of the Input-output linkages following 5 of OLS for the IV sample and the IV strategy are statistically significantly different from one another. The results correspond to comparing the coefficients and standard errors of columns (2) and (3). The results of the paired t-test suggest that the observed coefficients on input linkages from IV and OLS are statistically significantly different.