

University of Kent
School of Economics Discussion Papers

**Industry Agglomeration in a Developing Economy:
Evidence from India**

Amrit Amirapu Malavika Thirumalai Ananthakrishnan
Alex Klein

August 2026

KDPE 2604



Industry Agglomeration in a Developing Economy: Evidence from India*

Amrit Amirapu[†]

Malavika Thirumalai Ananthakrishnan[‡]

Alex Klein[§]

July 2026

Abstract

This paper examines the role of agglomeration economies in shaping patterns of industrial coagglomeration in post-liberalisation India. Using granular, establishment-level data from a 2005 census of all non-farm establishments, we estimate the relative importance of two Marshallian channels of agglomeration economies – input-output linkages and labour market pooling – in determining which industries locate together. To address concerns regarding endogeneity and measurement error we employ instrumental variables based on U.S. industrial data. We find large effects for both Marshallian channels – substantially larger than those the previous literature has found for advanced economies. By contrast, natural advantages appear less influential, though this may reflect measurement challenges. We also find that the two channels operate at distinct geographic scales: input-output linkages are similarly influential at both the district level and at the level of individual towns and villages, while labour market pooling effects appear more localized, as they are strongest at the town/village level. This pattern holds both in the full sample and within manufacturing alone. Our study is among the first detailed empirical assessments of coagglomeration patterns in a developing country, thus offering insights into the economic forces that shape spatial inequalities in such contexts.

Keywords: coagglomeration; agglomeration economies; Marshallian externalities; input–output linkages; labour market pooling; India

JEL codes: R12; R32; R23; O14; O18

*We wish to thank Sam Asher, A V Chari, Ritam Chaurey, Rana Hasan, Yi Jiang, Anirban Mitra, and seminar participants at the University of Kent and ISI Delhi. All errors remain our own. The authors may be contacted at a.amirapu@kent.ac.uk, malavika.ananthakrishnan@apu.edu.in, and Alexander.Klein@sussex.ac.uk.

[†]Department of Economics, University of Kent

[‡]School of Arts and Science, Azim Premji University

[§]Department of Economics, University of Sussex

1 Introduction

Firms benefit from being close to one another: more densely 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 specific 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 stresses various market mechanisms which incentivise firms to locate near each other (e.g., [Ellison and Glaeser, 1999](#)). The seminal work of [Marshall \(1956\)](#) argues that labour market pooling and proximity to suppliers and customers can largely explain the formation of geographic 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 (also referred to as ‘input-output linkages’) captures the idea that locating in geographic clusters allows firms to economise on transport costs when purchasing inputs or delivering goods to downstream customers. Despite a large body of work on the topic¹ – both theoretical and empirical – credible estimates of the relative importance of these Marshallian-type externalities have emerged only gradually. Furthermore, most of the existing studies have been conducted for advanced economies and have focused almost exclusively on the manufacturing sector.

This paper addresses a simple but unanswered question: which Marshallian mechanisms drive industrial co-location in a large developing economy? Despite substantial evidence that agglomeration economies exist in countries like India ([Lall et al., 2004](#); [Chauvin et al., 2017](#); [Amirapu et al., 2019](#)), the relative importance of specific channels — input-output linkages versus labour market pooling versus natural advantages — has not been rigorously assessed at scale in a developing-country context. This gap matters because the dominant Marshallian channel determines what kinds of policies can credibly promote productive agglomeration: if IO linkages dominate, supply-chain infrastructure may be the priority; if labour pooling dominates, the binding constraint may be labour mobility and commuting costs.

We answer this question by estimating both channels simultaneously using the methodology developed by [Ellison et al. \(2010\)](#) – henceforth EGK –, combined with employment and establishment level data from a census of 42 million Indian non-farm enterprises (including both manufacturing and service sector enterprises) collected in 2005.² The data allow us to study these channels at fine levels of geographic (district or town/village) and industrial (4-digit NIC code) disaggregation.³ 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.

EGK’s methodology requires measuring and exploiting variation in patterns of industrial coagglomeration: i.e., the extent to which firms from *different* industries locate near each other. Studying coagglomeration patterns is informative about the drivers of agglomeration because

¹E.g., [Deichmann et al., 2005](#); [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](#)

²Our primary data set is the Fifth Economic Census of India, which we supplement with various years of the Annual Survey of Industries, the Employment-Unemployment Surveys of the National Sample Survey Organization, and several other sources.

³NIC (i.e., National Industry Classification) codes are commonly used by Indian statistical agencies and roughly correspond to ISIC classification codes.

the strength of each Marshallian channel varies across industry pairs: firms in different industries may coagglomerate because they share suppliers or customers (input linkages), because they draw on the same labour pools, or because they benefit from overlapping knowledge spillovers. EGK developed a methodology that exploits this variation to disentangle the relative importance of the different agglomeration forces, and it is this approach that we follow here.

Our results can be summarised as follows. First, Marshallian agglomeration economies dominate natural advantages, at least insofar as the latter can be measured accurately given our data constraints. Second, the relative importance of the two Marshallian channels depends on the geographic scale: input-output linkages are similarly influential at both the district level and at the town/village level, while labour market pooling effects appear strongest at the town/village level. Third, the strength of agglomeration economies depends on the level of industrial classification in the anticipated way: more detailed industrial classification better captures the sources of industrial agglomeration – emphasizing the limits to attempting such analysis using more aggregated data – on which most of the prior literature relies.

The main contribution of this paper is to provide the first systematic empirical assessment of the relative importance of two core Marshallian agglomeration mechanisms — labour market pooling and proximity to suppliers and customers — in a large developing economy using a full census of all non-farm establishments. This complements the large recent literature on agglomeration externalities in advanced economies ([Overman and Puga, 2010](#); [Miyauchi, 2024](#); [Ellison et al., 2010](#); [Kolko, 2010](#); [Hanlon and Miscio, 2017](#); [Diodato et al., 2018](#); [Steijn et al., 2022](#)), and adds to a growing body of studies set in developing countries that examine the associations between geographical concentration and various economic mechanisms (e.g., [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 Janeiro; [Howard et al., 2016](#) for Vietnam). Although some of the previous literature suggests that agglomeration effects may be larger in developing countries than in advanced economies – providing additional motivation for our study – [Grover et al. \(2023\)](#) argue that these large estimates may reflect methodological limitations – a concern which our IV strategy is designed to address.

Within the developing-country literature, the closest studies to ours are [Howard et al. \(2016\)](#) and [Mukim \(2015\)](#). [Howard et al. \(2016\)](#) applies a similar methodology to Vietnamese firm data and finds that technology spillovers — not IO linkages — are the primary driver of coagglomeration in that context. In contrast, we study a much larger economy, conduct our main analysis at the 4-digit industry level (rather than 2-digit), add the service sector (omitted by all prior coagglomeration studies – including [Howard et al., 2016](#) and [Mukim, 2015](#)), and provide IV estimates robust to measurement error — a particularly important concern in analyses of developing-country firm-level data. Our findings also differ, suggesting a much larger role for IO linkages over longer distances and labour market pooling over shorter distances.

[Mukim \(2015\)](#) 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, first, in its question: we explore coagglomeration patterns *between* industries in order to shed light on Marshallian externalities while [Mukim \(2015\)](#) studies coagglomeration between formal and informal establishments *within* industries. Second, we use data from the Economic Census (EC) to construct our coagglomeration index rather than the Annual Survey of Industries (ASI). This is important because, while the ASI is only representative at the state level, the EC – being a census – is representative at any level of geography, allowing us to study which economic forces dominate at different geographical scales, includ-

ing fairly narrow levels of geographic aggregation. Finally, as mentioned above, our analysis is conducted at the 4-digit industry level,⁴ our data include manufacturing *and service* sector establishments, and our IV identification strategy facilitates estimation of the causal effects of our economic mechanisms.

Last, our paper also contributes to the wider, growing literature on spatial development in India (Desmet et al., 2013; Ghani et al., 2014; Duranton et al., 2016; Chauvin et al., 2017; Rossi-Hansberg, 2019; Amirapu et al., 2019), as well as a growing literature on the effects of industrial policies on spatial concentration in developing economies (Ghani et al., 2016; Li and Sinha Roy, 2020; Asher and Novosad, 2020; Baum-Snow et al., 2020; Gechter and Tsivandis, 2023). Our findings advance our understanding of how Marshallian agglomeration externalities operate in economic conditions (e.g., limited labour mobility, underdeveloped infrastructure, weak state capacity and legal institutions, a large informal sector) that differ substantially from those examined in advanced-economy studies.

The rest of the paper proceeds as follows. Section 2 provides background 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 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

⁴This allows for more precise estimation of inter-industry spillovers than would be possible using coarser measures, which we show in the paper matters a great deal.

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 choice, as overcrowding in cities has increased the cost of living and of 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 (Ghani et al., 2016; Asher and Novosad, 2020; Anantakrishnan, 2026). Even as progress was underway in reducing the cost of transportation, many frictions 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 et al., 2016; Morris and Pandey, 2009). These constraints, combined with high urban congestion costs and limited labour mobility, 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 (2005 EC), the Annual Survey of Industries for 2004/5 (2005 ASI), and Rounds 55, 60 and 61 of the Employment and Unemployment Surveys⁵ from the National Sample Survey Organization (NSSO). The 2005 EC is a census of all non-farm establishments⁶ in the country, covering both formal *and informal* establishments in both manufacturing *and service* industries. 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 2005 EC is an enormous data gathering process with the goal of surveying 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). Covering so many units is quite an achievement for a large country with poor state capacity and limited financial resources, but it is difficult to do with perfect accuracy, particularly given the reliance on relatively untrained enumerators.⁷ For this reason, we use an identification strategy that explicitly aims to address potential concerns regarding measurement error.

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

⁶According to documentation from the 2013 Economic Census, an establishment is defined as "... 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."

⁷The 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.

3.1 Construction of the Coagglomeration Index

We measure industrial coagglomeration following [Ellison and Glaeser \(1997\)](#) – henceforth EG – using data from the 2005 EC. 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 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:

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

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.⁸ As Eq. 1 makes clear, the coagglomeration index is closely related to the covariance, across regions, of the employment shares of industries i and j : the numerator captures the extent to which the two industries are over- or under-represented in the same regions, while the denominator normalises the index so that it is comparable across different distributions of aggregate employment. The value of the EG coagglomeration index ranges from -1 to 1. A high positive value indicates that the two industries tend to concentrate in the same regions; values close to 0 imply that the industry pair is spatially dispersed; and negative values imply that the two 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 304 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.⁹

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” (NIC 3511) and “Sea and coastal water transport” (NIC 6110)) while others are less so (e.g. “Dressing and dyeing of fur” (NIC 1820) and “Publishing of books” (NIC 2211)). The latter may reflect genuine (but idiosyncratic) coagglomeration patterns or simply measurement error in the Economic Census, as described above. Since measurement error in an independent variable leads to estimates that are biased and inconsistent, we address this concern with an instrumental variables strategy ([Durbin, 1954](#); [Cameron and Trivedi, 2005](#)). Figure 1 maps the employment shares for NIC code 3591 (manufacture of motorcycles) and NIC code 3592 (manufacture of bicycles and invalid carriages) – the most highly coagglomerated industry pair in the dataset. Both industries concentrate in Ludhiana, which holds 25.5% of national employment in motorcycles and 67.7% in bicycles.

⁸When constructing industrial coagglomeration indices among manufacturing industries only, we define x_m as region m ’s share in total manufacturing employment only (i.e. those industries with NIC 2004 divisions 15–37).

⁹At 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 aggregation are provided in Appendix B.2.

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

Note: Coagglomeration is γ_{ij} , defined in Eq. (1) and computed across 606 districts using employment from the 2005 Economic Census. The ten pairs shown are the highest-scoring of all 46,056 pairs formed from 304 4-digit NIC-2004 industries. They lie in the extreme right tail of the distribution: the median pair scores -0.000005 and the 99th percentile is 0.022, against 0.172 for the pair at the top of the table.

3.2 Causes of Industrial Coagglomeration

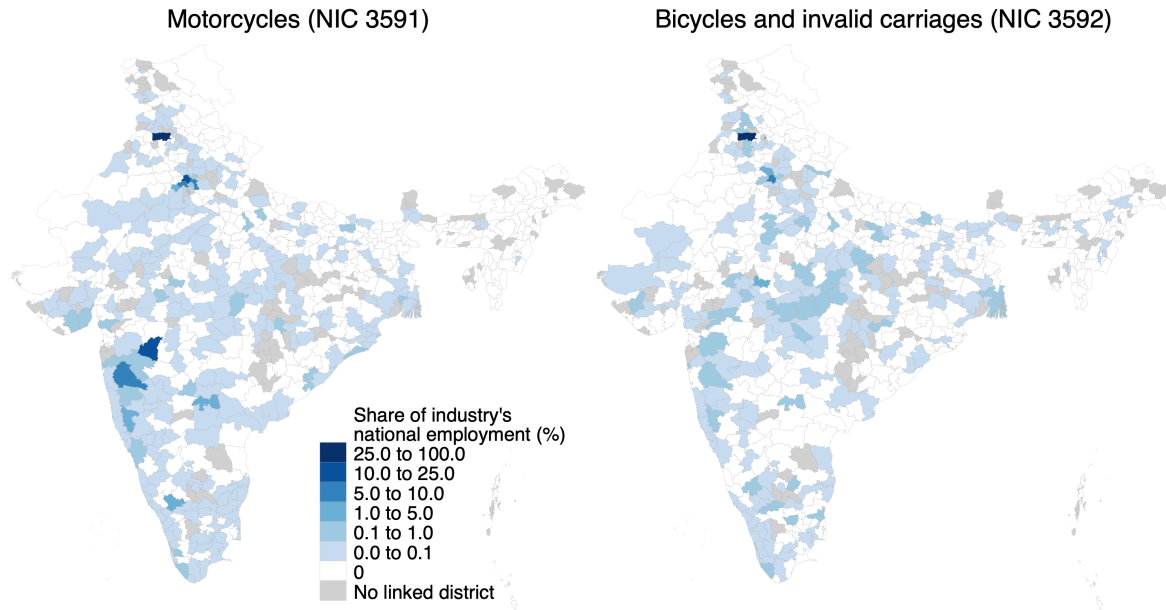
Industries may cluster in a given geographic space for a number of reasons. In this paper, we aim to separately identify the effects of two distinct sources of agglomeration economies: inter-industry input-output linkages and labour market pooling. To do so, we must also separately account for the effect 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.

3.2.1 Input-Output Linkages

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 : $\text{Input}_{i \leftarrow j}$. 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: $\text{Input}_{ij} = \max(\text{Input}_{i \leftarrow j}, \text{Input}_{j \leftarrow i})$.

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-

Figure 1: Geographical Concentration of Economic Activity: Motorcycles (3591) and Bicycles (3592)



Note: Each panel shades districts by s_{mi} , the district's share of the industry's total all-India employment in the 2005 Economic Census. This pair of industries has the highest coagglomeration index at the 4-digit district level, $\gamma_{ij} = 0.1717$. Ludhiana holds 25.5% of national employment in motorcycles and 67.7% in bicycles; Delhi holds a further 15.6% and 7.2%. Districts shaded white have no recorded employment in the industry. Areas shaded grey cannot be matched to a census district. Districts absent from the shapefile are omitted altogether; they hold 4.7% of national employment in motorcycles and 0.7% in bicycles. Both panels use the same class breaks, so a given shade denotes the same share in each.

by-commodity pairs into industry-by-industry pairs by mapping the 5-digit ASICC codes to 4-digit 2004 NIC codes using an official concordance. The mapping is not one-to-one, 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.

One caveat is important to note here. The ASI is representative of the organised manufacturing sector only. Since our input-output linkages are constructed from it, we are able to measure input linkages within manufacturing and between manufacturing and services (at least in one direction), but not within services.

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 regionally¹⁰ representative household surveys that include information on individuals' principal economic activities, including the occupations and in-

¹⁰NSS 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".

dustries 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 : $share_{io}$.¹¹ We identify individuals engaged in 458 different occupations (identified at NOC 3-digit level). Following 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.

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. 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 geographic 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) on five industry- and region-specific cost advantage factors, along with two district-level geographic features: river length (RL_m) and coastline (CL_m).

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

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 i 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.¹² Three of these five factors describe labour markets rather than physical resources. In this we are again following Ellison and Glaeser (1999) who define 'natural advantage' to include both "resource and labor-market natural advantages" (emphasis added).

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

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

We then use this term as a control variable in our main analysis to account for observed

¹¹When 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.

¹²Details on data construction are provided in Appendix A.1.

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	2005 ASI	Electricity Use	2005 ASI
2	Coal Price	2005 ASI	Coal Use	2005 ASI
3	Avg Manuf Wage	2005 ASI	Labour Use	2005 ASI
4	Share Tertiary Edu	2004-05 NSSO	Supervisory and Managerial Staff	2005 ASI
5	Share without degree	2004-05 NSSO	Unskilled and Contract Workers	2005 ASI

patterns in coagglomeration that can be explained through natural advantage alone.¹³ For comparison, EG show that roughly 20% of US manufacturing agglomeration can be explained by local natural advantages. One limitation of our approach is that we are able to measure only a limited set of natural cost advantages (the five factors above, plus the two geographic features), so our natural advantage index will not capture coagglomeration driven by advantages outside this set. Estimates of the role of natural advantages should therefore be interpreted as a lower bound.

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: measurement error and endogeneity. Measurement error is often a concern when using extremely large datasets from developing countries, where state capacity is limited – along with the resources available for enumerator training and data validation. This applies to the ASI, the NSS, and especially the EC datasets used here. Endogeneity is 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, the observed relationship between the Marshallian channels and the extent of coagglomeration would be driven by the initial choice to coagglomerate rather than the other way around.

Our empirical strategy is designed to address both concerns by instrumenting our measures of labour pooling and input-output linkages with their counterparts constructed from data for 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 also help with measurement error to the extent that (i) the data quality of the US instruments is higher than that of the original Indian measures, and (ii) 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:

$$\gamma_{ij} = \alpha + \beta_{NA}\gamma_{ij}^{CNA} + \beta_L\widehat{LabourCorr}_{ij} + \beta_{IO}\widehat{Input}_{ij} + \epsilon_{ij} \quad (4)$$

Here, γ_{ij} represents the pairwise coagglomeration index value for industries i and j ; γ_{ij}^{CNA}

¹³The cost-advantage factors underlying the natural advantage index (e.g. input prices and factor intensities from the ASI) are available only at the district level. We therefore construct γ_{ij}^{CNA} at the district level throughout, and use this same district-level control in the town/village specifications; a town/village-level natural advantage index is not feasible with the available data.

is the pairwise coagglomeration index for industries i and j predicted by the industries' exposure to natural advantages alone; $\widehat{Input}_{ij}$ and $\widehat{LabourCorr}_{ij}$ are the predicted values of the two main Marshallian channels (input-output linkages and labour market pooling) using the US instruments described in the next subsection.

Three features of our approach to inference deserve mention. First, because the unit of observation is an industry pair, we cluster standard errors in all specifications at the industry-pair level one step of aggregation above the level of analysis: at the three-digit pair level in the 4-digit specifications, and at the NIC Section-pair level (Sections A–Q) in the 2-digit specifications reported in the appendix. Second, our model is exactly identified (two instruments for two endogenous variables), and conventional t -based inference for IV can be unreliable when instruments are only moderately strong; we therefore report weak-instrument-robust Anderson–Rubin statistics — which remain valid regardless of instrument strength — alongside the 2SLS estimates in every IV specification (Andrews et al., 2019). Third, the natural advantages control γ_{ij}^{CNA} is a generated regressor: it is constructed from the estimated coefficients of equation 2, whose sampling uncertainty conventional standard errors ignore (Pagan, 1984). In Appendix Table 16 we show that our results are robust to propagating this uncertainty through a bootstrap of the full two-step procedure: the correction leaves the input-output estimates essentially unchanged and moderately widens the labour pooling standard errors, without altering any conclusion.

3.3.1 US Instrument Construction

Data on input-output linkages for US industries (i.e. $Input_{ij}^{US}$) are obtained from US Input-Output Tables for 2007.¹⁴ The relevant table identifies 405 industries using the 2007 NAICS (North American Industry Classification System) codes at the 6-digit level. Since there is no direct concordance between (Indian) NIC and (US) NAICS codes, we constructed a mapping using a series of intermediate concordances: we first mapped the 2007 NAICS codes to 2002 NAICS 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 are 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 NAICS level across 1090 occupations. To map the 2012 NAICS codes to 4-digit 2004 NIC codes, we first mapped the 2012 NAICS codes to 2007 NAICS 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: $LabourCorr_{ij}^{US}$.

3.4 Summary Statistics

Table 3 reports summary statistics for the main variables of interest for both the full OLS sample and the IV subsample. The full OLS sample covers 9,854 industry pairs for which the coagglomeration index, natural advantages control, and our Indian measures of IO linkages and labour correlation are all observed. The IV subsample (2,258 pairs) covers the sample of

¹⁴We 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.

industry pairs that could be matched to both US instruments and is considerably smaller because the series of concordances required to map US industry codes to Indian NIC codes yields a limited set of matched pairs. Because the IV subsample is determined by data availability rather than being drawn at random, we compare it with the remaining OLS-only observations ($N = 7,596$) to check whether there are significant differences. The final column of Table 3 reports the normalised difference in means between the two groups (that is, the difference in means divided by the pooled standard deviation).¹⁵ All four values lie between 0.06 and 0.16. In other words, the sample means differ by less than one sixth of a standard deviation, suggesting that the IV sample is broadly representative of the full sample in at least some important respects.¹⁶

Table 3: Summary Statistics (4-Digit NIC Code, District Level)

	Full Sample		IV Sample		Norm. diff.
	Mean	Std. Dev.	Mean	Std. Dev.	
EG Coagg. Index	0.0010	0.0060	0.0013	0.0044	0.058
Nat. Adv. Coagg. Index	0.0003	0.0002	0.0004	0.0002	0.155
Input-Output	0.0087	0.0517	0.0118	0.0578	0.074
Labour Correlation	0.1503	0.5221	0.1891	0.4347	0.102
<i>IV instruments (IV sample only):</i>					
Input-Output (US)			0.0146	0.0430	
Labour Correlation (US)			0.8436	0.0906	
Observations	9,854		2,258		

Note: This table reports means and standard deviations of the analysis variables, separately for the full OLS sample ($N = 9,854$) and for the subsample on which the IV estimates are identified ($N = 2,258$). The final column reports the normalised difference in means between the IV sample and the remaining OLS-only observations ($N = 7,596$): the difference in means divided by the pooled standard deviation of the two groups. The US instruments are defined only on the IV sample, so no full-sample moments or normalised differences are reported for them.

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 and the town/village level. In the Appendix, we present results at other levels of aggregation (i.e. 2-digit industry at district and town/village).

The main results are presented in Table 4, 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

¹⁵We report this rather than a t-statistic because at these sample sizes even negligible differences are statistically significant, whereas the normalised difference measures how large a difference is relative to the variation in the data.

¹⁶Note also that all normalised differences are well under 0.25, the conventional benchmark suggested by Imbens and Rubin (2015).

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

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.068*** (0.017)	0.077*** (0.024)	0.295*** (0.098)
Labour Correlation	-0.031** (0.013)	-0.041** (0.021)	0.441 (0.270)
Natural Advantages	0.021* (0.011)	0.042*** (0.014)	0.026 (0.018)
Sample	Full	IV	IV
R^2	0.005	0.019	
Observations	9,854	2,258	2,258
AR p-val (joint)			0.000
AR 95% CI: IO			[0.094, 0.743]
AR 95% CI: Labour			[-0.201, 1.253]

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. Standard errors, clustered at the three-digit industry-pair level, are reported 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 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.

of input-output linkages and natural advantages are positive but modest, while the coefficient on labour pooling is slightly negative. Our primary specification – shown in column 3 – uses the US-based instrumental variables to address the potential problems of endogeneity and measurement error discussed above. The results from our main specification (column 3 of Table 4) 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 for an industry pair is associated with an increase in observed coagglomeration (at the 4 digit industry x district level) of .30 standard deviations. For labour pooling, the coefficient (.44) is larger but noisier – just missing significance at the 10% level.¹⁷

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. It should be noted that direct comparison of standardised coefficients requires the standard deviations of the underlying measures to be similar across studies. For input-output linkages, the standard deviations are comparable (0.052 in our Indian sample versus 0.043 for the corresponding US instrument). For labour correlation, however, the Indian sample standard deviation (0.52) is substantially larger than the corresponding US measure (0.09), meaning that one standard deviation in our Indian labour correlation represents a considerably larger absolute increase in occupational overlap than in EGK’s US sample. The standardised coefficient comparison for labour pooling should therefore be interpreted with this caveat in mind. By contrast, the coefficient on our measure of natural advantages is small (.03) and

¹⁷The bootstrapped standard errors – employed to address the generated-regressor problem – reported in Appendix Table 16 imply a somewhat more conservative assessment still, while the effect is estimated much more sharply at the town/village level below.

insignificant – especially in comparison to the large effect found by EGK (.16) – though, as discussed above, this could be due to limitations in our ability to measure natural advantages with high accuracy and precision.

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 because there are many industry pairs for which we do not have data on the US instruments (see Section 3.3.1 above). To see whether the differences observed in column 3 are due to the instruments or to the change in sample, we reproduce the OLS results on the restricted IV sample in column 2. Reassuringly, the estimates in columns 1 and 2 are very similar, 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.¹⁸

In Table 5 we present the results from our first-stage regressions. Each instrument is a strong predictor of its Indian counterpart, and the Kleibergen-Paap rk Wald F statistic (18.5) is comfortably above the conventional rule-of-thumb threshold of 10. Because the Kleibergen-Paap statistic is a single system-wide measure, it cannot by itself establish that *each* of our two endogenous variables is separately identified: this would fail if, for example, both instruments predicted the same endogenous regressor. We therefore also report the Sanderson-Windmeijer conditional F statistic for each endogenous variable (Sanderson and Windmeijer, 2016), which tests exactly this. Both are moderately large in our main specification (51.3 for input-output linkages and 37.1 for labour pooling), confirming that the two channels are separately identified. Because conventional t-ratio-based inference can nonetheless be unreliable when instruments are only moderately strong (Lee et al., 2022), Table 4 also reports weak-instrument robust statistics: the joint Anderson-Rubin (AR) test rejects the null hypothesis that both endogenous coefficients are zero ($p < 0.001$), and the projection-based AR 95% confidence intervals for each endogenous variable are consistent with the reported 2SLS estimates. As for the exclusion restriction, we argue that because the instruments reflect industry relationships in another country, they should 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 6 presents results when constructing the coagglomeration index at an even finer geographic level: the town/village level.¹⁹ 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 be much greater – and more precisely estimated – at this finer level: a one standard deviation increase in our measure of labour pooling is associated with an increase of one standard deviation in observed coagglomeration (at the 4-digit industry x town level), suggesting that labour markets are very important in determining industrial location decisions but are also geographically constrained 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.

¹⁸A Wu-Hausman endogeneity test yields an F-statistic of 9.186 (with a corresponding p-value of 0.0001), further supporting our use of IV.

¹⁹Results from estimating specification 4 at coarser levels of industrial aggregation (i.e. 2-digit NIC code) are provided in the appendix (Tables 14 and 15). Briefly, the coefficients 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. These 2-digit estimates should, however, be treated with caution: the instruments are considerably weaker at this level of aggregation (the Sanderson-Windmeijer conditional F statistic for input-output linkages is 9.1, and the Kleibergen-Paap statistic is 3.3), and the Anderson-Rubin confidence intervals – which aim to account for this problem – are arbitrarily large.

Table 5: First-Stage of IV (4-digit NIC codes)

	(1) Input-Output	(2) Labour Correlation
Input-Output (US)	0.635*** (0.115)	-0.034* (0.014)
Labour Correlation (US)	0.044 (0.040)	0.141*** (0.023)
Natural Advantages	0.030 (0.022)	-0.028 (0.015)
Adjusted R^2	0.259	0.013
SW conditional F-Statistic	51.320	37.067
KP F-Statistic	18.526	18.526
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 are clustered at the three-digit industry-pair level. The Sanderson-Windmeijer conditional F statistic tests the strength of identification of the endogenous variable in that column, conditional on the other endogenous variable; the Kleibergen-Paap rk Wald F statistic is a single system-wide measure and is therefore identical across columns. 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.

One caveat is that the data, when constructed at the town/village level, systematically exclude several large cities²⁰ - 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-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 heavily on data from the ASI and is thus less likely to be able to explain coagglomeration patterns between industries in the service sector. For this reason, we re-estimate our main specification (equation 4) using only data from the manufacturing sector, at both of the levels of geographic aggregation considered above.²¹ The district-level results are presented in Table 7 and the town/village results in Table 8.

At the district level, the effect of input-output linkages is similar to what was found in Table 4. The coefficient on the labour pooling measure, however, is not statistically significant – and, in fact, the first stage of the IV estimation for labour pooling (Table 12 in the Appendix) shows that this instrument is weak when using only data from the manufacturing sector: the Sanderson-Windmeijer conditional F statistic for labour pooling falls to 9.3, below the con-

²⁰The town/village level data are drawn from the Socioeconomic High-resolution Rural-Urban Geographic dataset for India (SHRUG), a publicly available, harmonised panel database covering all rural and urban geographic units in India between 1990 and 2020 (Asher et al., 2021). SHRUG integrates data from India’s Economic Census, Population Census, and other administrative sources, harmonised across census years using a consistent set of sub-district geographic boundaries. A consequence of how SHRUG constructs consistent town and village boundaries across census rounds is that several large cities are excluded from the dataset, as their administrative boundaries changed substantially between census years.

²¹The coagglomeration index and the natural advantage index are both recomputed on the manufacturing employment universe, so that x_m is region m ’s share of manufacturing employment rather than of total employment.

Table 6: Main results (4-digit NIC code and town/village level)

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.048** (0.019)	0.049** (0.024)	0.217*** (0.063)
Labour Correlation	-0.015 (0.016)	-0.007 (0.028)	0.999*** (0.304)
Natural Advantages	0.041*** (0.011)	0.053*** (0.016)	0.040* (0.023)
Sample	Full	IV	IV
R^2	0.003	0.007	
Observations	9,854	2,258	2,258
AR p-val (joint)			0.000
AR 95% CI: IO			[0.107, 0.587]
AR 95% CI: Labour			[0.372, 2.011]

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. Standard errors, clustered at the three-digit industry-pair level, are reported 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 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.

ventional threshold of 10, while the corresponding statistic for input-output linkages remains high at 46.9. We therefore rely on the weak-instrument robust Anderson-Rubin confidence sets reported at the foot of the table: the set for input-output linkages remains informative, while the set for labour pooling is wide and uninformative.

The town/village results tell a different story. Input-output linkages remain positive and precisely estimated, with an AR confidence set of [0.063, 0.494]. Labour pooling, by contrast, is now large and positive, and its AR confidence set – [0.783, 11.533] – excludes zero. Two features of this comparison deserve emphasis. First, the first stage is common to the two specifications: they use the same 1,465 industry pairs and the same pair-level instruments, and the natural advantage control is measured at the district level in both cases, so the statistics reported in Table 12 apply to each of them. The labour pooling instrument is therefore equally weak in both, and weak identification cannot account for the difference between them; the only thing that changes across the two tables is the geographic scale at which the dependent variable is measured. Second, precisely because that instrument is weak, the magnitude of the town/village labour pooling estimate should not be read literally – the AR set extends to 11.5 – and we treat the informative content of the result as the sign and the exclusion of zero rather than the point estimate itself.

Taken together, these results reinforce the geographic pattern found in the full sample: input-output linkages operate at both larger (i.e. district) and smaller (i.e. town/village) scales, whereas labour market pooling appears strongest over shorter distances (i.e. within towns/villages). The absence of a clear labour pooling effect for manufacturing at the district level may thus reflect the limited spatial reach of the mechanism rather than its absence from the manufacturing sector. The natural advantage term remains small and statistically insignificant throughout.

Table 7: Manufacturing Sector (4-digit NIC code and district level)

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.065*** (0.016)	0.084*** (0.024)	0.214*** (0.080)
Labour Correlation	-0.045*** (0.015)	-0.055** (0.026)	-0.157 (0.672)
Natural Advantages	-0.005 (0.022)	0.037 (0.038)	0.028 (0.058)
Sample	Full	IV	IV
R^2	0.006	0.026	
Observations	7,099	1,465	1,465
AR p-val (joint)			0.025
AR 95% CI: IO			[0.023, 0.635]
AR 95% CI: Labour			[-3.029, 2.714]

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 manufacturing industries only (NIC 2004 divisions 15–37). Both the coagglomeration index and the natural advantage index are computed on the manufacturing employment universe. Standard errors, clustered at the three-digit industry-pair level, are reported 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 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. The Anderson-Rubin confidence sets are weak-instrument robust and are the preferred basis for inference on the two endogenous regressors; see Table 12.

5 Conclusion

Our study adds to the literature on agglomeration economies in several respects. This is one of the first papers to estimate the *causal* importance of specific Marshallian mechanisms in a large developing economy, using a census of both manufacturing and service establishments with IV estimation to address endogeneity and measurement error. We find that Marshallian agglomeration economies — particularly IO linkages and labour market pooling — are quantitatively important drivers of industrial coagglomeration in India, and that their relative importance varies systematically across geographic scales: IO linkages dominate at the district level, while labour pooling is more important at finer scales (towns and villages). The same contrast holds when the sample is restricted to manufacturing, where the estimation sample and the instruments are held fixed and only the geographic scale of the dependent variable changes, which suggests that the pattern reflects the spatial reach of the two mechanisms rather than differences in sectoral composition. This pattern is consistent with high labour transport costs constraining the geographic reach of labour market pooling in India relative to advanced economies.

Three caveats are important. First, our analysis is limited to a single cross-section (2005), and we cannot rule out that some co-location patterns reflect historical industrial policy (e.g., licensing decisions from the pre-liberalisation period) rather than market-driven agglomeration. Second, our IO linkage measure is constructed from the ASI, which covers formal manufacturing only; our estimates for industry pairs involving the service sector are therefore likely to be attenuated. Third, our IV sample covers only a subset of industry pairs for which concordances

Table 8: Manufacturing Sector (4-digit NIC code and town/village level)

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.059*** (0.017)	0.065*** (0.024)	0.198*** (0.057)
Labour Correlation	-0.024 (0.016)	-0.016 (0.028)	1.882** (0.771)
Natural Advantages	-0.001 (0.021)	0.067** (0.029)	-0.108 (0.098)
Sample	Full	IV	IV
R^2	0.004	0.016	
Observations	7,099	1,465	1,465
AR p-val (joint)			0.000
AR 95% CI: IO			[0.063, 0.494]
AR 95% CI: Labour			[0.783, 11.533]

Note: This table repeats the specification of Table 7 with the coagglomeration index measured across towns and villages rather than districts, for manufacturing industries only (NIC 2004 divisions 15–37). The estimation sample, the instruments and the natural advantage control are identical to those in Table 7; only the geographic scale at which the dependent variable is constructed differs, and the first stage reported in Table 12 therefore applies to both tables. Standard errors, clustered at the three-digit industry-pair level, are reported 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 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. Because the labour pooling instrument is weak in this sample (Sanderson-Windmeijer conditional F of 9.3), inference on the two endogenous regressors should rest on the Anderson-Rubin confidence sets rather than on the point estimates and conventional standard errors.

to US data are available; while the IV sample is close to the broader OLS sample on observable dimensions (Section 3.4), unobservable differences may exist and may bias the results. Addressing these limitations — through panel variation using other Economic Census waves, improved natural advantage measures, or natural experiments — represents a natural extension of this research.

For policy, our results suggest that agglomeration forces in India may not be limited to the well-documented manufacturing clusters of the post-liberalisation era, but may extend to services — which have been the primary engine of India’s growth since liberalisation (Fan et al., 2023). Policies that improve inter-industry linkages (e.g., industrial parks, supply-chain infrastructure) may generate coagglomeration spillovers beyond the manufacturing sector. Similarly, the finding that labour pooling effects are geographically concentrated at small spatial scales suggests that reducing the costs of short-distance worker mobility — through improved housing, local transport infrastructure, urban planning, or even changing social norms²² — could enhance the productivity gains from agglomeration in a developing-country context where such costs remain high.

²²For example, Amirapu et al. (2022) argue that gender-based social norms increase the costs of worker mobility in Bangladesh.

References

- Alkay, Elif and Geoffrey JD Hewings**, “The determinants of agglomeration for the manufacturing sector in the Istanbul metropolitan area,” *The Annals of Regional Science*, 2012, 48, 225–245.
- Amirapu, Amrit and Arvind Subramanian**, “Manufacturing or Services? An Indian Illustration of a Development Dilemma,” Working Paper 408, Center for Global Development 2015.
- , **M. Niaz Asadullah, and Zaki Wahhaj**, “Social barriers to female migration: Theory and evidence from Bangladesh,” *Journal of Development Economics*, 2022, 158, 102891.
- , **Rana Hasan, Yi Jiang, and Alex Klein**, “Geographic concentration in Indian manufacturing and service industries: Evidence from 1998 to 2013,” *Asian Economic Policy Review*, 2019, 14 (1), 148–168.
- Anantakrishnan, Malavika Thirumalai**, “Rural Roads and Their Impact on Urban Centres: Evidence from Indian Cities and Towns,” 2026. Working paper.
- Andrews, Isaiah, James H Stock, and Liyang Sun**, “Weak instruments in instrumental variables regression: Theory and practice,” *Annual Review of Economics*, 2019, 11, 727–753.
- Asher, Sam and Paul Novosad**, “Rural roads and local economic development,” *American Economic Review*, 2020, 110 (3), 797–823.
- , **Tobias Lunt, Ryu Matsuura, and Paul Novosad**, “Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the shrug open data platform,” *The World Bank Economic Review*, 2021, 35 (4), 845–871.
- Barufi, Ana Maria Bonomi, Eduardo Amaral Haddad, and Peter Nijkamp**, “Industrial scope of agglomeration economies in Brazil,” *The Annals of Regional Science*, 2016, 56, 707–755.
- Baum-Snow, Nathaniel, J Vernon Henderson, Matthew A Turner, Qinghua Zhang, and Loren Brandt**, “Does investment in national highways help or hurt hinterland city growth?,” *Journal of Urban Economics*, 2020, 115, 103124.
- Cameron, A. Colin and Pravin K. Trivedi**, *Microeconometrics: Methods and Applications*, Cambridge University Press, 2005.
- Chauvin, Juan Pablo, Edward Glaeser, Yueran Ma, and Kristina Tobio**, “What is different about urbanization in rich and poor countries? Cities in Brazil, China, India and the United States,” *Journal of Urban Economics*, 2017, 98, 17–49.
- Deichmann, Uwe, Kai Kaiser, Somik V. Lall, and Zmarak Shalizi**, “Agglomeration, transport, and regional development in Indonesia,” 2005. Working paper.
- Desmet, Klaus, Ejaz Ghani, and Stephen O’Connell**, “India’s Spatial Development,” Economic Premise 124, The World Bank 2013.
- Diodato, Dario, Frank Neffke, and Neave O’Clery**, “Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time,” *Journal of Urban Economics*, 2018, 106, 1–26.

- Duranton, Gilles, Syed Ejaz Ghani, Arti Grover Goswami, and William Robert Kerr**, “A detailed anatomy of factor misallocation in India,” Policy Research Working Paper 7547, The World Bank 2016.
- Durbin, J.**, “Errors in Variables,” *Revue de l’Institut International de Statistique / Review of the International Statistical Institute*, 1954, 22 (1/3), 23–32.
- Dwivedi, Aasheerwad and Amaresh Dubey**, “Coagglomeration And New Economic Activity: Evidence From Formal And Informal Manufacturing Firms,” *Journal of Developmental Entrepreneurship*, 2022, 27 (02), 2250013.
- Ellison, Glenn and Edward L Glaeser**, “Geographic concentration in US manufacturing industries: a dartboard approach,” *Journal of Political Economy*, 1997, 105 (5), 889–927.
- **and** —, “The geographic concentration of industry: does natural advantage explain agglomeration?,” *American Economic Review*, 1999, 89 (2), 311–316.
- , —, **and William R Kerr**, “What causes industry agglomeration? Evidence from coagglomeration patterns,” *American Economic Review*, 2010, 100 (3), 1195–1213.
- Fan, Tianyu, Michael Peters, and Fabrizio Zilibotti**, “Growing Like India—the Unequal Effects of Service-Led Growth,” *Econometrica*, 2023, 91 (4), 1457–1494.
- Fernandes, Ana and Gunjan Sharma**, “The role of infrastructure, governance, education and industrial policy,” Working Paper C-35005-INC-1, International Growth Centre 2012.
- Gechter, Michael and Nick Tsivandis**, “Spatial Spillovers from High-Rise Developments: Evidence from the Mumbai Mills,” 2023. Working paper.
- Ghani, Ejaz, Arti Grover Goswami, and William R. Kerr**, “Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing,” *The Economic Journal*, April 2016, 126 (591), 317–357.
- , **William R. Kerr, and Stephen O’Connell**, “Spatial determinants of entrepreneurship in India,” *Regional Studies*, 2014, 48 (6), 1071–1089.
- Grover, Arti, Somik Lall, and Jonathan Timmis**, “Agglomeration economies in developing countries: A meta-analysis,” *Regional Science and Urban Economics*, 2023, 101, 103901.
- Hanlon, W Walker and Antonio Miscio**, “Agglomeration: A long-run panel data approach,” *Journal of Urban Economics*, 2017, 99, 1–14.
- He, Canfei and Junsong Wang**, “Geographical agglomeration and co-agglomeration of foreign and domestic enterprises: A case study of Chinese manufacturing industries,” *Post-Communist Economies*, 2010, 22 (3), 323–343.
- Howard, Emma, Carol Newman, and Finn Tarp**, “Measuring industry coagglomeration and identifying the driving forces,” *Journal of Economic Geography*, 2016, 16 (5), 1055–1078.
- Imbens, Guido W. and Donald B. Rubin**, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press, 2015.
- Kathuria, Vinish**, “What Causes Agglomeration—Policy or Infrastructure? A Study of Indian Organised Manufacturing,” *Economic and Political Weekly*, 2016, pp. 33–44.

- Kochhar, Kalpana, Utsav Kumar, Raghuram Rajan, Arvind Subramanian, and Ioannis Tokatlidis**, “India’s pattern of development: What happened, what follows?,” *Journal of Monetary Economics*, July 2006, 53 (5), 981–1019.
- Kolko, Jed**, “Urbanization, agglomeration, and coagglomeration of service industries,” in “Agglomeration economics,” University of Chicago Press, 2010, pp. 151–180.
- Kone, Zovanga L, Maggie Y Liu, Aaditya Mattoo, Caglar Ozden, and Siddharth Sharma**, “Internal borders and migration in India,” *Journal of Economic Geography*, 2018, 18 (4), 729–759.
- Lall, Somik V, Zmarak Shalizi, and Uwe Deichmann**, “Agglomeration economies and productivity in Indian industry,” *Journal of Development Economics*, 2004, 73 (2), 643–673.
- Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter**, “Valid t-Ratio Inference for IV,” *American Economic Review*, 2022, 112 (10), 3260–3290.
- Li, Yue and Sutirtha Sinha Roy**, “The Employment Effect of Place-Based Policies,” Policy Research Working Paper 9477, The World Bank 2020.
- Lu, Jiangyong and Zhigang Tao**, “Trends and determinants of China’s industrial agglomeration,” *Journal of Urban Economics*, 2009, 65 (2), 167–180.
- Marshall, Alfred**, *Principles of Economics*, 8th ed., London: Macmillan, 1956.
- MIT**, “MIT Geodata Repository,” <https://geodata.mit.edu> n.d.
- Miyauchi, Yuhei**, “Matching and Agglomeration: Theory and Evidence From Japanese Firm-to-Firm Trade,” *Econometrica*, 2024, 92 (6), 1869–1905.
- Morris, Sebastian and Ajay Pandey**, “Land Markets in India: Distortions and Issues,” in Nirmal Mohanty, Runa Sarkar, and Ajay Pandey, eds., *India Infrastructure Report 2009: Land — A Critical Resource for Infrastructure*, New Delhi: Oxford University Press, 2009.
- Mukim, Megha**, “Coagglomeration of formal and informal industry: evidence from India,” *Journal of Economic Geography*, 2015, 15 (2), 329–351.
- Overman, Henry G and Diego Puga**, “Labor pooling as a source of agglomeration: An empirical investigation,” in “Agglomeration economics,” University of Chicago Press, 2010, pp. 133–150.
- Pagan, Adrian**, “Econometric issues in the analysis of regressions with generated regressors,” *International Economic Review*, 1984, 25 (1), 221–247.
- Resende, Marcelo**, “Industrial coagglomeration: some state-level evidence for Brazil,” *Nova Economia*, 2015, 25, 181–194.
- Rodrik, Dani and Arvind Subramanian**, “From “Hindu Growth” to Productivity Surge: The Mystery of the Indian Growth Transition,” *IMF Staff Papers*, 2005, 52 (2), 193–228.
- Rossi-Hansberg, Esteban**, “Geography of growth and development,” in “Oxford Research Encyclopedia of Economics and Finance” 2019.

Sanderson, Eleanor and Frank Windmeijer, “A weak instrument F -test in linear IV models with multiple endogenous variables,” *Journal of Econometrics*, 2016, 190 (2), 212–221.

Steijn, Mathieu P.A., Hans R.A. Koster, and Frank G. Van Oort, “The dynamics of industry agglomeration: Evidence from 44 years of coagglomeration patterns,” *Journal of Urban Economics*, 2022, 130.

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 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 0150	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 0150	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 codes)

	(1) Input-Output	(2) Labour Correlation
Input-Output (US)	0.704*** (0.129)	-0.029* (0.014)
Labour Correlation (US)	0.051 (0.076)	0.113** (0.037)
Natural Advantages	0.085* (0.038)	0.084* (0.033)
Adjusted R^2	0.355	0.010
SW conditional F-Statistic	46.895	9.333
KP F-Statistic	4.655	4.655
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. This first stage is common to Tables 7 and 8: both are estimated on the same industry pairs, both instrument with the same pair-level US measures, and both use the district-level natural advantage control, so the estimates and test statistics below are identical for the two specifications and are reported once. Standard errors in parentheses are clustered at the three-digit industry-pair level. The Sanderson-Windmeijer conditional F statistic tests the strength of identification of the endogenous variable in that column, conditional on the other endogenous variable; the Kleibergen-Paap rk Wald F statistic is a single system-wide measure and is therefore identical across columns. Note that the conditional F statistic for labour pooling (9.3) falls below the conventional threshold of 10 in this manufacturing-only sample, which is why the labour pooling estimates in Tables 7 and 8 should be assessed using the weak-instrument robust Anderson-Rubin confidence sets reported there. 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 codes)

	(1) Input-Output	(2) Labour Correlation
Input-Output (US)	0.327 (0.169)	-0.093* (0.034)
Labour Correlation (US)	0.013 (0.036)	0.352*** (0.084)
Natural Advantages	0.039 (0.051)	0.011 (0.084)
Adjusted R^2	0.115	0.136
SW conditional F-Statistic	9.126	20.789
KP F-Statistic	3.256	3.256
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 NIC Section-pair level (Sections A–Q). The Sanderson-Windmeijer conditional F statistic tests the strength of identification of the endogenous variable in that column, conditional on the other endogenous variable; the Kleibergen-Paap rk Wald F statistic is a single system-wide measure and is therefore identical across columns. Both instruments are weaker at this coarser level of industrial aggregation than in the 4-digit specifications, so the accompanying Anderson-Rubin statistics are the appropriate basis for inference. 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 (0.088)	0.215* (0.124)	1.278** (0.506)
Labour Correlation	-0.045 (0.039)	-0.135* (0.074)	0.084 (0.260)
Natural Advantages	-0.012 (0.082)	0.066 (0.121)	-0.042 (0.128)
Sample R^2	Full 0.012	IV 0.035	IV
Observations	649	428	428
AR p-val (joint)			0.058
AR 95% CI: IO			entire grid
AR 95% CI: Labour			entire grid

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. Standard errors, clustered at the NIC Section-pair level (Sections A–Q, the level of aggregation above 2-digit Divisions), are reported 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 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.

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

	EG Coagglomeration Index		
	(1) OLS	(2) OLS	(3) IV
Input-Output	0.059 (0.051)	0.146* (0.083)	0.928 (0.578)
Labour Correlation	-0.046 (0.087)	-0.148* (0.086)	0.012 (0.227)
Natural Advantages	0.076 (0.074)	0.128 (0.104)	0.048 (0.115)
Sample	Full	IV	IV
R^2	0.009	0.038	
Observations	649	428	428
AR p-val (joint)			0.063
AR 95% CI: IO			entire grid
AR 95% CI: Labour			entire grid

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. Standard errors, clustered at the NIC Section-pair level (Sections A–Q, the level of aggregation above 2-digit Divisions), are reported 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 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.

Table 16: Robustness to the generated-regressor problem: bootstrap standard errors

	Coefficient	Analytic SE	Bootstrap SE (nonparametric)	Bootstrap SE (parametric)
<i>Panel A: IV, district level (4-digit NIC)</i>				
Input-Output	0.295	(0.098)	(0.105)	(0.109)
Labour Correlation	0.441	(0.270)	(0.300)	(0.358)
Natural Advantages	0.025	(0.017)	(0.094)	(0.040)
<i>Panel B: IV, town/village level (4-digit NIC)</i>				
Input-Output	0.217	(0.063)	(0.076)	(0.079)
Labour Correlation	0.999	(0.304)	(0.342)	(0.379)
Natural Advantages	0.039	(0.022)	(0.093)	(0.039)
<i>Panel C: OLS, full sample, district level</i>				
Input-Output	0.068	(0.017)	(0.018)	(0.018)
Labour Correlation	-0.031	(0.013)	(0.013)	(0.013)
Observations (IV / OLS)		2,258 / 9,854		
Bootstrap replications		1,000		

Note: This table examines the sensitivity of our main estimates to the generated-regressor problem: the natural advantages control γ^{CNA} is constructed from the estimated coefficients of equation 2, and conventional standard errors treat those coefficients as known (Pagan, 1984). The first column reproduces the point estimates; the second reports the analytic standard errors used in the main tables (clustered at the three-digit industry-pair level). The last two columns report bootstrap standard errors from 1,000 replications of the full two-step procedure, in which the natural-advantage prediction equation is re-estimated and γ^{CNA} rebuilt in every replication before a cluster resample of the industry pairs: in the “nonparametric” column the prediction equation is re-estimated on a resample of its industries; in the “parametric” column its coefficients are drawn from their estimated two-way (district and industry) cluster-robust sampling distribution. The correction leaves the OLS estimates and the IV input-output coefficients essentially unaffected and moderately widens the IV labour pooling standard errors; no conclusion in the paper changes. The larger nonparametric standard error on the natural advantages coefficient reflects a small number of extreme industry resamples (the bootstrap distribution has heavy tails for this coefficient only).