Overcoming the Agglomeration Paradox: Skill-Dependent FDI and Urbanization in China

Samantha A. Vortherms

Appendix

Measuring Access to Urban Services

China's population data is of notorious poor quality and open to both manipulation and misinterpretation (Chan 2007). In a country with high levels of quantification to meet political goals (Ghosh 2020; Wallace 2022), local governments manipulate population numbers to reach political ends. In some measures, population numbers are inflated to reach nationally-defined urbanization goals. In other measures, population numbers use restricted numbers to improve government performance indicators measured per capita. Other times, data are not reported at all by lower level governments to obfuscate policy reform. This analysis took these concerns seriously by validating population data across multiple sources, with specific care to ensure continuity of population data both across municipal units and across time.

Types of population data used, their definitions, and their sources:

- *Hukou* population. The *hukou* population includes the number of people holding a municipality's *hukou*, regardless of *hukou* type (urban/rural/unified resident). The *hukou* population is tracked by the Public Security Bureau in each municipality, based on official records. This is different than the "long-term resident population 常住人口" which is measured based on surveys. Long term residents include those registered locally in the *hukou* system and live there, people whose *hukou* is registered in some other location but have lived there for more than six months, and those in the local hukou system who have left town for less than six months (Li and Li 2016). I use *hukou* population as the key indicator of the local population because this is the group with full, permanent access to local government services. While long-term resident populations may have access to some government services, this access is often limited—such as non-local students attending grade school in Shanghai but not being allowed to take the college entrance exam—and can be rescinded.

- Non-agricultural population. The non-agricultural population 非人口are defined as individuals with non-agricultural local *hukou*. These are the residents who have access

to urban government services, compared with rural government services for "agricultural" *hukou* holders. Colloquially, this group is often referred to as "urban" citizens or the urban population. This designation is strictly bureaucratic: not all non-agricultural *hukou* holders live and work in urban settings, but instead inherited their non-agricultural status from their parents or were transferred to non-agricultural status through a local naturalization process.

- Urban population 城人口. The urban population has three different definitions, depending on municipality and time. The urban population can refer to residents living in urban districts under the municipal government (市). This accounting excludes individuals living in small urban centers in rural counties or townships () outside of the core urban areas. This definition was most common before 2000. The second definition is any individual living in urban areas, including both municipal centers and rural townships and is usually estimated using sampled surveys. The third definition is anyone with non-agricultural *hukou*. After the national urbanization reform in 2014 that pushed the development of a "new type" of *hukou* system that aligned non-agricultural with urban definitions, many cities simply re-labeled their non-agricultural populations (非人口) "urban" (城人口).

Estimating Urban Transfers

Urban transfers are estimated as the net number of trans-municipal migrants granted urban *hukou*. Each year, the official *hukou* population grows by natural increase of births minus deaths and by mechanical growth—in-migration minus out-migration.[1] Using the official *hukou* population, we can estimate the net number of migrants as the annual change in *hukou* population minus the natural change in *hukou* population. Any growth, or decline, in the *hukou* population not capture by births and deaths is due to net migration. Net *hukou* transfers, then, was calculated as the annual change in *hukou* population not attributed to natural growth.

Estimating the Urban Benefits Population

There are two types of migrants in Chinese municipalities: those that cross municipal borders (non-local migrants) and those that move from the countryside to the urban center crossing county lines (local migrants). Both of these populations face barriers to accessing urban government services. *Hukou* transfers, because it is based on the aggregate *hukou* population of the entire municipality, only captures the migration across municipal lines. Local migrants who move from the countryside to the urban center are not captured in this measure.

Before 2014, local rural-to-urban migrants could be identified systematically because municipalities broke down *hukou* population numbers by agricultural and non-agricultural. After 2014, in response to a national reform, municipalities stopped reporting the number of agricultural and non-agricultural *hukou* holders, instead switching to estimates of the urban resident population. In 2010, 315 municipalities reported the breakdown of *hukou*

type of the population. By 2015, only 178 municipalities reported the numbers. In some cases, agricultural and non-agricultural *hukous* were integrated, but in most cases, *hukou* differences remain but are simply not reported.

When the breakdown of agricultural and non-agricultural *hukou* populations were reported, the true population with access to urban services (non-agricultural *hukou* holders) could be measured. Without this breakdown, however, I rely on an estimate of the urban benefits population. I take urban resident population, which includes registered urban *hukou* holders and migrants who have lived in urban districts for at least six months and subtract the estimated migrant population of the city. When non-agricultural *hukou* holders live outside of urban districts, this measure is an underestimate of the urban benefits population or when migrants live in rural counties instead of urban districts.

Determining Skill Level

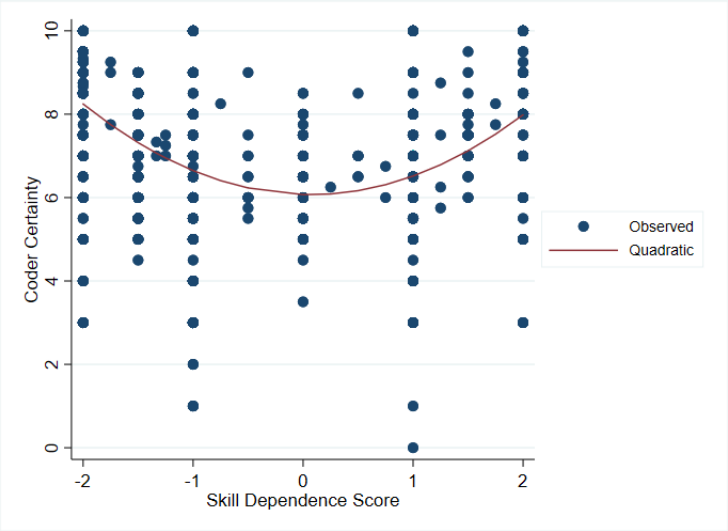
The key independent variable for the analysis is dependence on skilled labor. To measure skill dependence, I use content analysis and a random forest model to identify skill dependence. Random forest models are a form of supervised machine learning models used to classify observations. An algorithmic random forest approach has been shown to outperform logistic regression for predicting outcomes out of sample (Muchlinski *et al.* 2016). The random forest used a hand-coded training dataset of 4,400 firms, approximately one percent of the total sample.

Hand-coding firms

Ideally, I would be able to use more objective measures, such as employee average education or degree/certification attainment, to measure skill dependence. Without these data, however, I used firm activities as an indicator of the likelihood firm productions are dependent on more high-skilled workers relative to low-skilled workers. With the assistance of four research assistants, we coded each firm based on a four-point likelihood scale for skill dependence. All firms are likely to require some high-skilled labor—factories need technicians and engineers as well as line workers. I conceptualized the independent variable as skill dependence rather than skill use to capture the relative reliance on skilled workers. Factories, while requiring skilled workers, high disproportionately more low-skilled workers. Similarly, technology innovation firms also require workers without "high-skill" training to function, but relatively speaking, the function of the firm depends more on high-skilled labor than the low-skilled labor hired.

Research assistants were instructed to identify a firm as high-skilled dependent if the majority of operations needed to be conducted by someone with a post-secondary degree (associates or higher) or a skilled technician.[2] Business operations such as "research and design," for example, were flagged as higher-skilled activities while "warehousing" ranked low on the skill dependence. Firms expected to hire significantly more low-skilled workers than high-skilled workers rank lower on the skill dependency spectrum.

Figure A1: Non-linear Relationship between Skill Coding and Coder Certainty



For each coded firm, research assistants were also asked to rate the certainty in their coding on a scale from 0 to 10, with 10 being absolutely certain and 0 being a complete guess. The average certainty score was 7.4 across all firms. The certainty score was correlated with skill dependence coding, with both high skilled and low-skilled firms having higher certainty scores (average 8.2) than those in the middle ranges (average 6.6) (Figure A1). This suggests that skill ratings closer to zero had higher coder uncertainty. Because the initial coding used 0 as a midpoint, weighting by certainty biases scores towards zero in the middle range of the skill dependence rating.

To improve coding quality, the research team met weekly to identify and discuss uncertain cases. Each team member presented one case whose coding they were certain on and one case whose coding they were uncertain on. The group then discussed the uncertain cases to come to an agreement about the proper coding. These discussions helped improve reliability of ratings within the group, adding to deeper discussions of what dependence on high skill meant in practice. Additionally, approximately 15 percent of the sample was double coded with at least two coders coding the same firm. Overall, there was a 75 percent agreement rate across firms. When collapsed into high skilled and low skilled categories, the coders had an 84 percent agreement rate across firms. Table A1 presents pairwise intercoder reliability score, calculated by Cohen's κ, which takes into account the possibility of coding similarities due to error.

Coders were also asked to code the industry classifications from the Industrial Classification for National Economic Activity. Together the four coders rated the approximately 1,400 industrial categories—also known as four digit industrial coding—on a three point scale from -1 for low skilled to 1 for high skilled. Most manufacturing industrial categories fall into the low-skilled range, except for highly specialized manufacturing, such as the spacecraft and launch vehicle manufacturing (industry C3742). Other high-skilled industries include

Table A1: Cohen's Kappa, Pairwise between coders

Coders	1	2	3	4
1				
2	0.33			
3	0.40	0.36		
4	0.46	0.51	0.47	

management and financial firms, like the economic affairs management industry (industry S9225). The Industrial Classification for National Economic Activity include notes on many industries that coders used to classify on the three point scale.

Through this process, we also built a dictionary of high-skilled words based on the description of the business operations. The dictionary of high-skilled words was then rated. Each word or phrase was ranked on a 0 to 3 scale, where 0 cleaned out words that were too ambiguous to signal skill dependence and 3 signified skill dependence with high certainty. I then ran a weighted content analysis on business descriptions to identify how many high-skilled words were included.

Random forest model

The outcome variable of the random forest model was the continuous skill dependence rating discussed in the previous section. The random forest model used a total of ten predictor variables: skill-dependence content analysis, industry rating, registered capital, founding year, joint venture status, investor country of origin (OECD or non-OECD), and geographic location within China (three regions and urban/rural). I use the *rforest* package for Stata to implement the random forest model (Schonlau and Zou 2020). As a robustness check, I used logistic regression to predict skill dependence as an alternative to the random forest model. The validation error in the logit regression was on average 40 percent, suggesting the random forest technique vastly outperforms a logistic alternative in this case (results not shown).

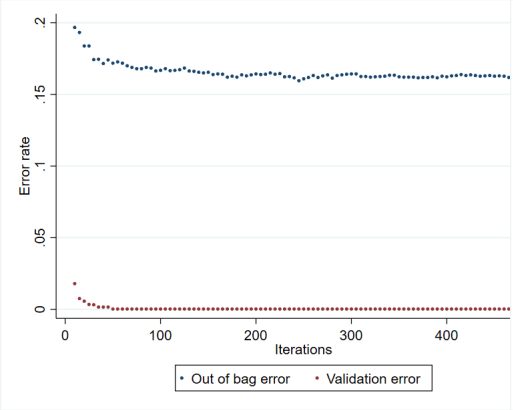
Figure A2 presents diagnostic tests for the number of iterations and number of included variables. Out-of-bag error minimizes and stabilizes around 300 iterations (0.159) and four variables (0.158). Validation error minimizes and stabilizes around 75 iterations (0.0002) and constant across number of variables (0.0002). The final model included 4 variables and 350 iterations.

Figure A3 identifies the relative importance of the variables used in the random forest model. The three most important predictors are registered capital, the high-skilled content analysis of business operations, and founding year.

Skill-dependent firms were defined as the top quartile of firms rated by the random forest model. I collapse the continuous skill-dependent variable into this indicator variable to compensate for the potential error and uncertainty in the hand coding of firms. An alternative would be to include skill-dependence as a continuous variable, which would be more

Figure A2: Random Forest Diagnostic Tests

(a) Iteration Test



(b) Variable inclusion test

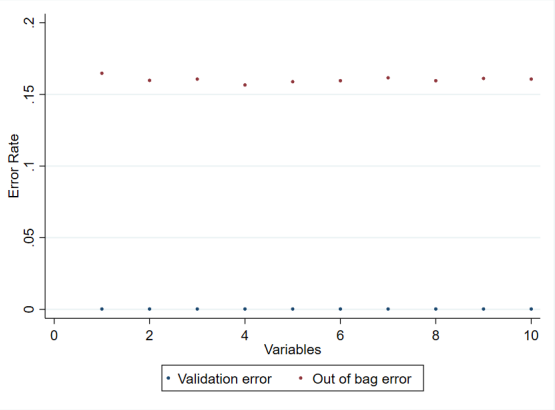
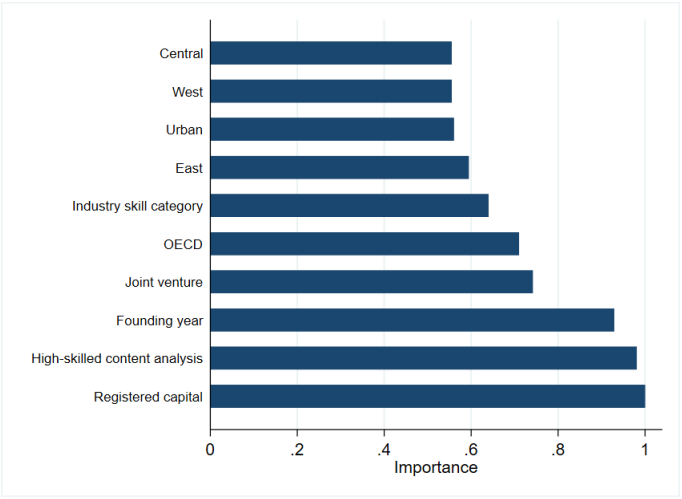


Figure A3: Variable Importance



justifiable with objective measures of skill dependence. Because human coding introduces noise, an indicator variable reduces the influence of the noisy data generating process.

Results Tables

Table A2: Descriptive Statistics

Variable	Obs.	Mean	Std. Dev.	Min.	Max.
Hukou Transfers (net)	1,420	-6,095	122,760	-1,497,327	1,648,589
Urban Benefits Population	1,166	1,648,220	1,653,522	-7,542,705	14,200,000
Skill-dependent foreign capital (per total)	1,414	0.28	0.23	0.00	0.99
Skill-dependent established firms (cap. Per total)	1,414	0.27	0.22	0.00	0.99
Established firms (cap. per total)	1,414	0.89	0.14	0.01	1.00
Foreign Capital (log)	1,414	12.90	1.94	6.26	19.66
Exports (log)	1,420	11.71	2.05	4.82	17.85
GDP (log)	1,420	7.50	0.92	5.25	10.55
Primary Industry (per GDP)	1,420	0.12	0.08	0.00	0.48
Science Spending (per local spending)	1,420	0.02	0.02	0.00	0.17
Local Expenditures (log)	1,420	5.87	0.71	3.22	9.03
Registered Population	1,420	4,548,598	3,136,156	202,500	34,200,000
Migrant stock (est.)	1,420	979,503	1,594,805	56,137	12,800,000
Villages upgraded (any)	1,420	0.19	0.39	0.00	1.00

Table A3: Fixed-effects models of registered foreign capital on *hukou* transfers and the size of the urban-benefits population

VARIABLES	(1) Hukou Transfers	(2) Hukou Transfers	(3) Hukou Transfers	(4) Hukou Transfers
High-Skill Dependence		84,161**		
		(37,142)		
Cap. in Est. High-Skill			73,164**	71,015**
(per total)			(35,065)	(35,112)
Established Capital				-28,484
(per total)				(25,148)
Reg. Foreign Capital (log)	-12,049*	-13,554*	-13,485*	-15,280**
	(7,299)	(7,316)	(7,321)	(7,490)
Exports (log)	-1,112	-2,042	-1,861	-1,915
	(6,859)	(6,858)	(6,858)	(6,857)
Municipal GDP (log)	-18,172	-22,417	-23,261	-23,956
	(45,652)	(45,607)	(45,649)	(45,647)
Primary Sector (per GDP)	500,192	475,578	478,809	478,215
	(360,420)	(359,920)	(360,027)	(359,982)
Science Spending	632,795	633,024	631,898	628,105
	(413,721)	(412,959)	(413,103)	(413,064)
Gov't Expenditures	-69,480	-69,501	-68,342	-67,330
	(43,270)	(43,191)	(43,209)	(43,213)
Registered Population	-0.00404	-0.00402	-0.00419	-0.00410
	(0.00350)	(0.00350)	(0.00350)	(0.00350)
Migrant Stock	0.0957*	0.0941*	0.0929*	0.0924*
	(0.0557)	(0.0556)	(0.0556)	(0.0556)
Upgraded Villages (any)	-15,749*	-16,328*	-16,180*	-16,321*
	(9,485)	(9,471)	(9,473)	(9,473)
Lagged DV	-0.0972***	-0.0993***	-0.0983***	-0.0988***
	(0.0279)	(0.0279)	(0.0279)	(0.0279)
Constant	535,224	580,963	584,401	631,946*
	(372,843)	(372,703)	(373,031)	(375,339)
Observations	1,420	1,420	1,420	1,420
R-squared	0.029	0.034	0.033	0.034
Number of Municipalities	287	287	287	287

All models include municipal and year fixed effects. Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A4: Fixed-effects models of registered foreign capital on *hukou* transfers and the size of the urban-benefits population

VARIABLES	(1) UBP	(2) UBP	(3) UBP	(4) UBP
High-Skill Dependence		100,991**		
		(46,876)		
Cap. in Est. High-Skill			109,404**	108,004**
(per total)			(45,163)	(45,590)
Established Capital				-7,638
(per total)				(32,963)
Reg. Foreign Capital (log)	-16,659*	-19,753**	-20,176**	-20,404**
	(9,381)	(9,471)	(9,467)	(9,523)
Exports (log)	-1,893	-3,426	-4,184	-4,163
	(9,446)	(9,453)	(9,467)	(9,473)
Municipal GDP (log)	159,326**	149,872**	148,941**	148,400**
	(66,443)	(66,450)	(66,397)	(66,474)
Primary Sector (per GDP)	656,356	639,506	630,862	631,401
	(447,051)	(446,189)	(445,933)	(446,181)
Science Spending	134,370	133,128	132,477	129,811
	(452,736)	(451,794)	(451,478)	(451,870)
Gov't Expenditures	-93,635*	-97,980*	-96,196*	-95,906*
	(55,588)	(55,509)	(55,443)	(55,487)
Registered Population	0.00386	0.00378	0.00357	0.00360
	(0.00413)	(0.00412)	(0.00412)	(0.00412)
Migrant Stock	0.327***	0.325***	0.322***	0.322***
	(0.0617)	(0.0616)	(0.0616)	(0.0616)
Upgraded Villages (any)	-11,350	-11,485	-11,282	-11,283
	(11,202)	(11,179)	(11,171)	(11,177)
Lagged DV	0.802***	0.804***	0.804***	0.804***
	(0.0374)	(0.0373)	(0.0373)	(0.0373)
Constant	-477,447	-347,346	-329,008	-317,164
	(546,489)	(548,685)	(548,404)	(551,078)
Observations	1,124	1,124	1,124	1,124
R-squared	0.574	0.577	0.577	0.577
Number of Municipalities	238	238	238	238

All models include municipal and year fixed effects. Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A5: Fixed-effects models of registered foreign capital on *hukou* transfers by region

VARIABLES	(1) East	(2) Central/West	(3) East	(4) Central/West
High-Skill Dependence	219,315***	-26,079		
	(33,204)	(55,876)		
Established High-Skill (per total)			201,740***	-23,146
			(32,780)	(51,903)
Established Capital (per total)			-39,592	-5,398
			(25,488)	(36,302)
Reg. Foreign Capital (log)	-23,121***	-1,741	-23,286***	-2,439
	(5,443)	(12,267)	(5,430)	(13,035)
Exports (log)	-2,435	-6,090	-2,563	-6,067
	(6,413)	(10,329)	(6,406)	(10,340)
Municipal GDP (log)	-71,163*	-67,965	-66,653*	-66,722
	(40,015)	(73,654)	(39,847)	(73,752)
Primary Sector (per GDP)	-304,102	512,610	-193,190	517,971
	(386,172)	(503,611)	(384,649)	(504,103)
Science Spending	282,148	1.054e+06	267,473	1.053e+06
	(291,443)	(722,862)	(290,802)	(723,473)
Gov't Expenditures	71,322**	-183,355***	82,152**	-182,663***
	(36,144)	(67,182)	(36,019)	(67,228)
Migrant Stock	0.0327	-0.173	0.0275	-0.171
	(0.0427)	(0.156)	(0.0427)	(0.156)
Registered Population	0.0567	-0.00172	0.0585	-0.00166
	(0.0395)	(0.00451)	(0.0394)	(0.00451)
Upgraded Villages (any)	-8,080	-23,271*	-6,974	-23,226*
	(8,233)	(14,062)	(8,206)	(14,073)
Lagged DV	0.173*	-0.108***	0.169*	-0.108***
	(0.0961)	(0.0343)	(0.0967)	(0.0343)
Constant	141,257	1.589e+06***	78,936	1.586e+06**
	(314,944)	(612,318)	(313,015)	(615,880)
Observations	568	852	568	852
R-squared	0.204	0.047	0.209	0.047
Number of Municipalities	114	173	114	173

All models include municipal and year fixed effects. Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A6: Fixed-effects models of registered foreign capital on the size of the urban-benefits population by region

VARIABLES	(1) East	(2) Central/West	(3) East	(4) Central/West
High-Skill Dependence	427,185***	50,328		
	(66,665)	(56,668)		
Established High-Skill (per total)			364,896***	58,235
			(65,102)	(54,792)
Established Capital (per total)			-86,921	26,110
			(58,064)	(36,857)
Reg. Foreign Capital (log)	-44,820***	-7,749	-43,341***	-6,081
	(10,427)	(13,370)	(10,487)	(13,655)
Exports (log)	-22,530	-837.2	-27,493	-1,531
	(16,611)	(10,651)	(16,718)	(10,695)
Municipal GDP (log)	-148,906	85,430	-124,724	86,524
	(105,097)	(78,727)	(104,908)	(78,794)
Primary Sector (per GDP)	846,729	668,602	969,370	656,912
	(875,998)	(494,921)	(880,963)	(495,134)
Science Spending	204,925	503,966	128,517	516,480
	(480,291)	(646,988)	(482,785)	(647,306)
Gov't Expenditures	-31,332	-73,662	7,287	-75,900
	(77,346)	(68,910)	(76,862)	(68,918)
Migrant Stock	-0.172**	1.750***	-0.175**	1.750***
	(0.0767)	(0.144)	(0.0769)	(0.144)
Registered Population	0.162**	-0.0103**	0.157**	-0.0105**
	(0.0751)	(0.00442)	(0.0755)	(0.00443)
Villages Upgraded (any)	467.4	-18,644	4,598	-19,015
	(16,025)	(13,083)	(16,086)	(13,090)
Lagged DV	0.570***	0.775***	0.572***	0.775***
	(0.0455)	(0.0532)	(0.0457)	(0.0532)
Constant	2.343e+06***	-888,846	2.084e+06**	-915,678
	(888,388)	(637,727)	(882,259)	(642,055)
Observations	420	704	420	704
R-squared	0.674	0.651	0.673	0.651
Number of Municipalities	89	149	89	149

All models include municipal and year fixed effects. Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

References

Chan, Kam Wing. 2007. Misconceptions and Complexities in the Study of China's Cities: Definitions, Statistics, and Implications. *Eurasian Geography and Economics*, **48**(4), 383–412.

Ghosh, Arunabh. 2020. *Making It Count: Statistics and Statecraft in the Early People's Republic of China*. Histories of Economic Life. Princeton, NJ: Princeton University Press.

Li, Xiru, and Li, Rui. 2016. 常住人口和流人口如何分*[How to distinguish between long-term residents and migrant population]*.

Muchlinski, David, Siroky, David, He, Jingrui, and Kocher, Matthew. 2016. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, **24**(1), 87–103.

Schonlau, Matthias, and Zou, Rosie Yuyan. 2020. The random forest algorithm for statistical learning. *The Stata Journal*, **20**(1), 3–29.

Wallace, Jeremy. 2022. *Seeking Truth and Hiding Facts:Information, Ideology, and Authoritarianism in China*. Oxford: Oxford University Press.