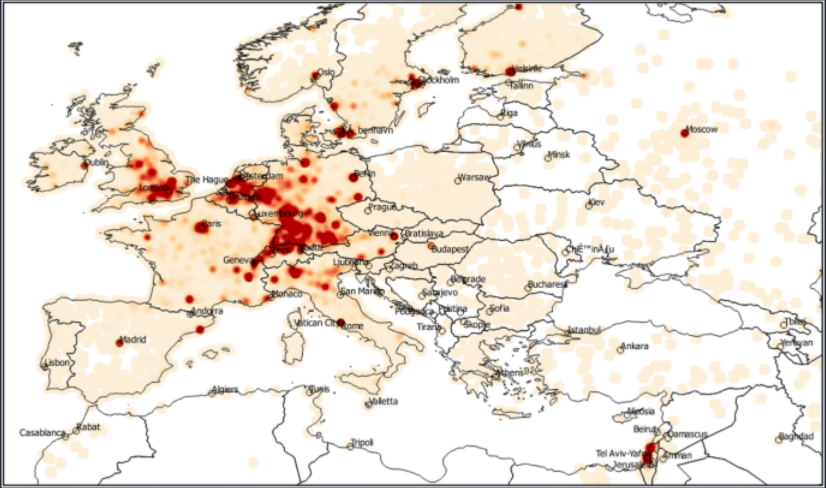
Spatial data

DSIER [/dɪˈzaɪər/] — Summer 2024

Julian Hinz

Bielefeld University

Patents' inventors location



Stek (2020)

Patterns in Spatial data

Why are inventors concentrated in certain locations?

1. common influences
   - quality of education, infrastructures, etc.
   - economic activity, jobs, etc.
2. spillover effects
   - social interactions
   - information sharing
   - skill hubs

Patterns in Spatial data

Why are inventors concentrated in certain locations?

1. common influences
   - quality of education, infrastructures, etc.
   - economic activity, jobs, etc.
2. spillover effects
   - social interactions
   - information sharing
   - skill hubs

Patterns in Spatial data

Why are inventors concentrated in certain locations?

1. common influences
   - quality of education, infrastructures, etc.
   - economic activity, jobs, etc.
2. spillover effects
   - social interactions
   - information sharing
   - skill hubs

Analysis of Spatial data

- Can you identify the two effects?
- Idea: Estimate the following model via OLS

$$y_{ig} = \gamma x_{ig} + \beta m_y(y_g) + \delta m_x(x_{ig}) + \epsilon_{ig} \tag{1}$$

- y_{ig} is the number of patents of inventor i geolocalised in area g
- x_{ig} are individual characteristics (age)
- m_y, m_x are aggregations of variables that are spatially connected with location g

  $\rightarrow$ e.g. avg. age and avg. number of patents in the same location

Reflection Problem

The average outcome for the group is an aggregation of outcomes or behaviours over other group members, i.e. aggregation of individual characteristics over other group members

$\rightarrow$ Multicollinearity

Roadmap

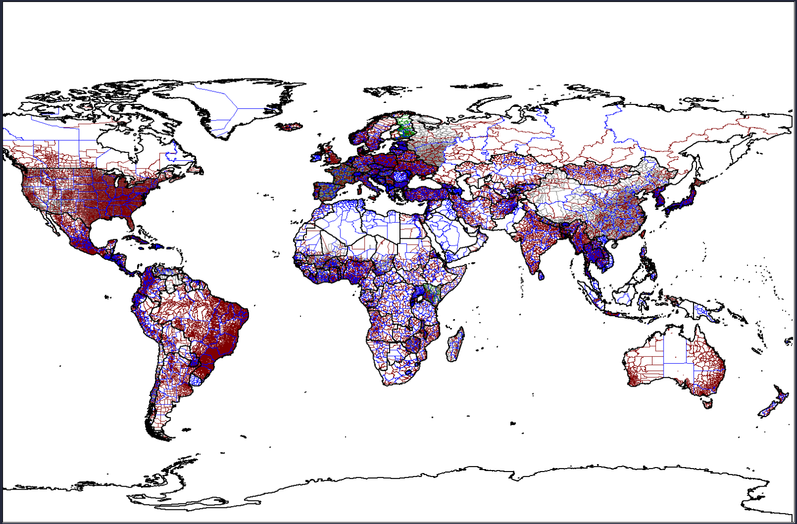
1. Terminology and concepts
    - terminology with Spatial Data
    - simple indexes of spatial concentration
2. Non-randomness in spatial data
3. Spatial Models
4. Identification of Spatial Models
5. Application:
    - Dreher et al. 2019 on China Foreign Assistance

TERMINOLOGY AND CONCEPTS

Content of Spatial data

- position data in 2D (or 3D) (location of inventors)
  $\rightarrow$ sometimes entities: polygons
- attribute data (number of patents)
- metadata related to the position data (characteristics of location)

Units of geographical space



Simple indicators of concentration

How to measure concentration of patents across regions in a country?

- Krugman specialization/concentration index
- Spatial Gini Index

Simple indicators of concentration

How to measure concentration of patents across regions in a country?

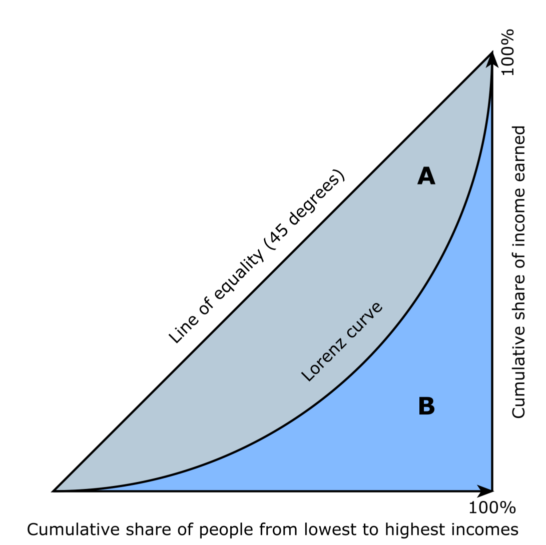
- Krugman specialization/concentration index

$$Conc = \sum_{g=1}^{n} |s_g - s| \tag{2}$$

where s_g is the number of patents per capita in region g with g=$\{1, \ldots, n\}$, while s is the number per capita in the whole economy

Gini Index

Rank people by income, instead of regions by number of patents



Gini Index

It's equivalent to the relative mean absolute difference

$$G = \frac{\sum_{g=1}^{n}\sum_{j=1}^{n}|x_g - x_j|}{2\sum_{g=1}^{n}\sum_{j=1}^{n}x_j} = \frac{\sum_{g=1}^{n}\sum_{j=1}^{n}|x_g - x_j|}{2n\sum_{j=1}^{n}x_j} = \frac{\sum_{g=1}^{n}\sum_{j=1}^{n}|x_g - x_j|}{2n^2\bar{x}} \tag{3}$$

where x_g is the number of patents in region g. The Gini is the mean absolute difference of all pairs of regions of the country divided by the average, $\bar{x}$, to normalize for scale.

Spatial decomposition of the Gini coefficient

Rey and Smith (2013) decompose the numerator as follows

$$\sum_{g=1}^{n}\sum_{j=1}^{n}|x_g - x_j| = \underbrace{\sum_{g=1}^{n}\sum_{j=1}^{n} w_{gj}\,|x_g - x_j|}_{a} + \underbrace{\sum_{g=1}^{n}\sum_{j=1}^{n}(1 - w_{gj})\,|x_g - x_j|}_{b} \tag{4}$$

where w_{gj} is binary spatial weights expressing the neighbor relationship between locations g and j.

- a<b positive spatial autocorrelation
- a>b negative spatial autocorrelation

Spatial decomposition of the Gini coefficient

Rey and Smith (2013) decompose the numerator as follows

$$\sum_{g=1}^{n}\sum_{j=1}^{n}|x_g - x_j| = \underbrace{\sum_{g=1}^{n}\sum_{j=1}^{n} w_{gj}\,|x_g - x_j|}_{a} + \underbrace{\sum_{g=1}^{n}\sum_{j=1}^{n}(1 - w_{gj})\,|x_g - x_j|}_{b} \tag{4}$$

where w_{gj} is binary spatial weights expressing the neighbor relationship between locations g and j.

- a<b positive spatial autocorrelation
- a>b negative spatial autocorrelation

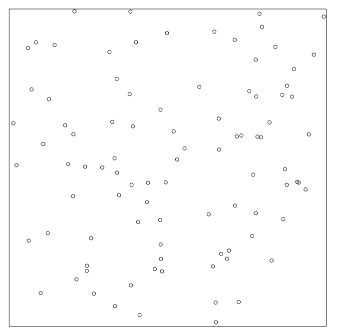
Computing Spatial Autocorrelation

The X,Y coordinates refer to the geometric centroids of the 325 Municipalities in Greece (Programme Kallikratis) in 2011.

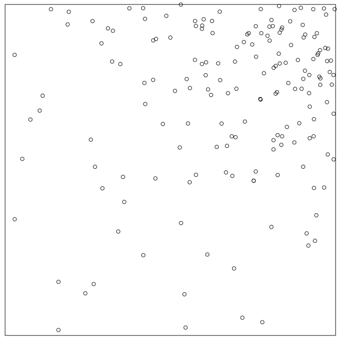
```
library(lctools)
data(GR.Municipalities)
names(GR.Municipalities@data)
 [1] "OBJECTID"   "X"          "Y"          "Name"       "CodeELSTAT" "PopM01"     "PopF01"     "PopTot01"   "UnemrM01"
[11] "UnemrT01"   "PrSect01"   "Foreig01"   "Income01"

myDF<-cbind(GR.Municipalities@data$Income01,GR.Municipalities@data$X, GR.Municipalities@data$Y)
myDF[!complete.cases(myDF),]
     [,1] [,2] [,3]
myDF.new<-na.omit(myDF)
bw<-12  # with 12 neighbours
wt<-'Binary'  # each neighbour is weighted with the same weight
spGini(myDF.new[,2:3],bw,myDF.new[,1],wt)
                     x
------------  --------
Gini           0.14550
gwGini         0.00342   # this is a of the slide before (the inequality among nearest (geographically) neighbours)
nsGini         0.14208   # this is b (the inequality among furthest (geographically) neighbours)
gwGini.frac    0.02351
nsGini.frac    0.97649
```

NON-RANDOMNESS IN SPATIAL DATA



Complete Random Allocation in 2D



Incomplete Random Allocation in 2D

Point Process

Poisson process:

- n points (denoted x1,x2,...,xn)
- randomly spatially distributed in a region W
    - each point xi are selected from a random uniform distributed over W
    - coordinates of each point location is independent
- n is randomly generated from a Poisson distribution with intensity (EV) λ

$$\Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{5}$$

The expected number of points to fall in a window with area |W| is λ|W|

Example of Point Processes

Generate a point pattern realization of the Poisson process

1. Homogenous (CRS): with λ=100 points/unit area in a 2D window with domain $x, y \in [0, 1]$

```
library(spatstat)
pois.pp <- rpoispp(lambda = 100, win = owin(c(0,1), c(0,1)))
plot(pois.pp)
```

2. Inhomogenous: with λ=f(x,y) points/unit area in a 2D window with domain $x, y \in [0, 1]$

```
lambda.u <- function(x,y){1000 * x^2 * y^2 + 50}
pois.inh.pp <- rpoispp(lambda = lambda.u, win = owin(c(0,1), c(0,1)))
plot(pois.inh.pp)
```

Relevance in Economics Literature

- Traditionally, indexes such as Gini, Krugman concentration are compared over time (e.g. Imbs and Wacziarg (2003))

- Detecting non-randomness is often non evident:
  - Ellison and Gleaser (1997): adjust index of spatial concentration for industrial concentration
  - See Combes and Overman (2004) for a discussion

The economics of spatial Non-randomness

1. random allocation, characteristics of location varies
   - farmers randomly allocated, but their crops depends on soil etc. (Holmes and Lee, 2012)
2. non-random allocation, location characteristics no causal effect on outcomes
   - R&D in Silicon Valley (Ellison and Gleaser, 1997)
3. random allocation, interactions matters
   - college dormitory allocation and peer effect in the choice of majors (Sacerdote, 2001)
4. non-random allocation, interactions matters
   - childhood neighborhood (Gibbons, 2013)

SPATIAL MODELS

Linear Spatial Model

$$y_{ig} = \beta m_y(y_g) + \delta m_x(x_{ig}) + \theta m_z(z_{ig}) + \sigma m_v(v_{ig}) + \epsilon_{ig} \tag{6}$$

where

- y_{ig} is the outcome of obs. i geolocalised in area g
- x_{ig} are individual characteristics
- z_{ig} are characteristics of other entities or object other than i
- v_{ig} are unobservable characteristics
- m(.) are aggregations of variables that are spatially connected with location g.
  - passive (externalities) vs deliberate (interactions)
  - pure technological externality or pecuniary externality

Specifying the interconnections

Are typically LC of the observations in neighbouring locations with group weights

$$m_x(x, s_i) = \sum_{j=1}^{M} g_{ij}(s_i, s_j)x_j = G_{xi}x \tag{7}$$

where G_{xi} is a $1 \times M$ row vector of the set of weights relating to location s_i, and x is an $M \times 1$ column vector of x for locations $s_1, s_2, \ldots, s_M$.

Matrix notation is more convenient for all observations i, where G is an $N \times M$ matrix, so

$$m_x(x, s) = G_x x \tag{8}$$

and similarly for z, y, and v.

Specifying the interconnections

Which type of neighborhood is represented in the following interaction structure?

$$
G = \begin{array}{cc} & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} & \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \end{array}, \quad GG = \begin{array}{cc} & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} & \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \end{array}. \qquad (3.7)
$$

Type of spatial models

If N=M we can rewrite (5) using the interconnection matrix

$$y = X\gamma + G_y y\beta + G_x X\delta + G_z Z\theta + G_v V\sigma + \epsilon \tag{9}$$

Spatial Econometrics literature usually treat $G_y = G_x = G_z$ (Spatial Lags)

Restrictions on this model

- Spatial Autoregressive model: $\delta = \theta = \sigma = 0$
- Spatially Lagged model: $\beta = \sigma = 0$
- Spatial Durbin Model: $\sigma = 0$
- Spatial Error Model:$\beta = \theta = 0$

Identification in spatial models

- reflection problem
- correlated unobservables or common shocks
- sorting - presence of OVs correlated with location decision

The reflection problem

Assume $G = G_y = G_x = G_z$ and $u = GV\sigma + \epsilon$ and rewrite (8)

$$y = X\gamma + Gy\beta + GX\delta + GZ\theta + u$$
$$Gy = GX\frac{(\gamma + \delta)}{(1-\beta)} + GZ\theta + u$$
$$y = X\tilde{\gamma} + GX\tilde{\delta} + GZ\tilde{\theta} + \tilde{u}$$

with $\tilde{\gamma} = \frac{\gamma}{(1-\beta)}$, $\tilde{\delta} = \frac{\gamma\beta+\delta}{(1-\beta)}$ and $\tilde{\theta} = \frac{\theta}{(1-\beta)}$.

Manski (1993) ”reflection problem”: β, δ and θ cannot be separately identified!

Solutions to the reflection problem

- Use of non-linear functional forms
  - Brock and Durlauf 2001 use binary outcome and estimate the probability of smoking
- imposing exclusion restrictions
  - $\beta = 0$ no endogenous effects
  - assume away GX, no contextual effects (Gaviria and Raphael, 2001)
- use incomplete interactions s.t. $GG \neq G$
  - non linearities in group membership: Calvo-Armengol et al. 2009, Liu and Lee 2010

Spatially Correlated Shocks

Identification is a problem whenever u=$G_v\sigma + \epsilon$ is correlated with x or z

1. group membership is exogeneous and correlation is due to OV
   - unobserved region characteristics that encourage inventors to patent (as human capital externalities in Moretti 2004)

2. group membership is endogenous and correlation is due to sorting
   - more innovative types move into areas with higher returns to innovation

Solution to Spatially Correlated Shocks: Spatial Differencing

Consists in transforming variables by subtracting spatial means (Holmes, 1998)

In our setting it means knowing G_v and then multiply by a transformation matrix $[I - G_v]$ to give:

$$y - G_v y = (X - G_v X)\gamma + (G_v - G_v G_x)X\delta + (G_z - G_v G_z)Z\theta + \zeta$$

If $plim(G_v - G_v G_v)v = 0$ then OLS is consistent

When is it possible? When $G_v G_v = G_v$, block structure as before

APPLICATION

Journal of Development Economics 140 (2019) 44–71

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

African leaders and the geography of China's foreign assistance

Axel Dreher [a,b,c,d,e], Andreas Fuchs [f,g], Roland Hodler [c,e,h], Bradley C. Parks [i,j], Paul A. Raschky [k,*], Michael J. Tierney [l]

[a] Alfred-Weber-Institute for Economics, Heidelberg University, Germany
[b] KOF Swiss Economic Institute, Switzerland
[c] CEPR, UK
[d] Georg-August University Goettingen, Germany
[e] CESifo, Germany
[f] Faculty of Economic and Social Sciences, Helmut-Schmidt-University Hamburg (HSU/UniBwH), Germany
[g] Research Area "Poverty Reduction, Equity, and Development", Kiel Institute for the World Economy, Germany
[h] Department of Economics and SIAW-HSG, University of St. Gallen, Switzerland
[i] AidData, Global Research Institute, The College of William and Mary, USA
[j] Center for Global Development, USA
[k] Department of Economics, Monash University, Australia
[l] Department of Government, The College of William and Mary, USA

ARTICLE INFO

JEL classifications:
D73
F35
P33
R11

Keywords:
foreign aid
favoritism
political capture

ABSTRACT

We investigate whether foreign aid from China is prone to political capture in aid-receiving countries. Specifically, we examine whether more Chinese aid is allocated to the birth regions of political leaders, controlling for indicators of need and various fixed effects. We collect data on 117 African leaders' birthplaces and geocode 1650 Chinese development projects across 2969 physical locations in Africa from 2000 to 2012. Our econometric results show that political leaders' birth regions receive substantially larger financial flows from China in the years when they hold power compared to what the same region receives at other times. We find evidence that these biases are a consequence of electoral competition: Chinese aid disproportionately benefits politically privileged regions in country-years when incumbents face upcoming elections and when electoral competitiveness is high. We observe no such pattern of favoritism in the spatial distribution of World Bank development projects.

Caption

Map of Chinese Aid Value

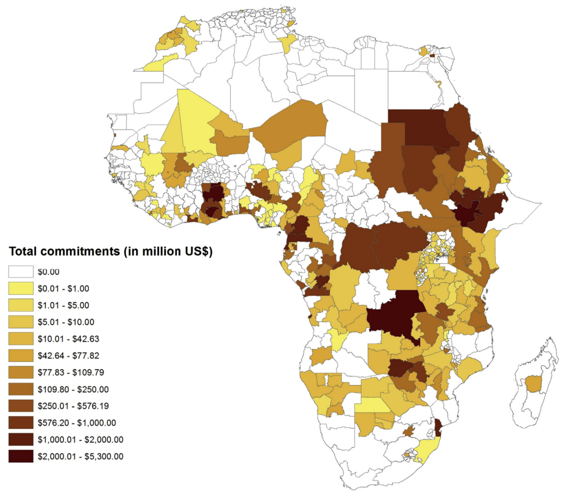


Fig. 1. Value of Chinese aid projects per subnational unit in Africa (total value in million 2009 US$, 2000–11, ADM1).

Map of Leaders’ birthplaces

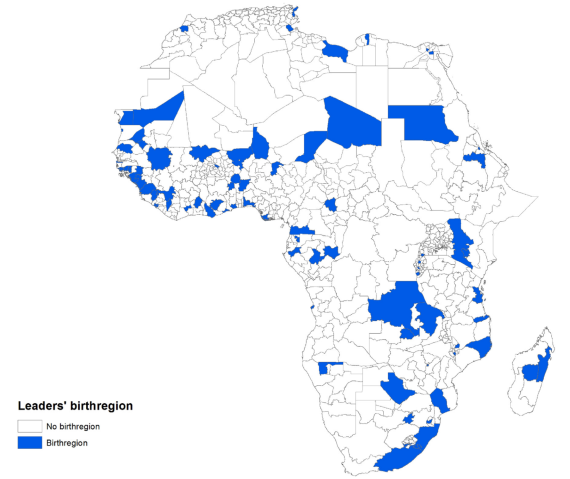


Fig. 2. Birth regions of effective political leaders in Africa (2000–11, ADM1).

Empirical strategy

Do current political leaders' birthplaces matter for the allocation of Chinese aid?

$$Aid_{ict} = \alpha + \gamma Birthregion_{ict} + \epsilon_{ict}$$

where $Birthregion_{ict}$ is equal to 1 if the political leader of country c in year t was born in administrative region i, and zero otherwise.

Problems?

They apply Spatial differencing:

$$Aid_{ict} = \alpha_{ct} + \delta_{ic} + \sum_j \beta_j X_{ic}^j + \gamma Birthregion_{ict} + \epsilon_{ict}$$

Table 2
Birth regions and China's aid I, ADM1, 2000–11.

	(1) Total OLS	(2) Total PPML	(3) ODA OLS	(4) ODA PPML	(5) Total OLS	(6) Total PPML	(7) ODA OLS	(8) ODA PPML
Birthregion	0.688**	0.969***	0.283	0.921	1.082**	0.267*	0.569**	2.257***
	(0.324)	(0.359)	(0.209)	(0.564)	(0.423)	(0.142)	(0.252)	(0.328)
Light2000 (in logs)	0.293**	0.218	0.242*	−0.117				
	(0.119)	(0.158)	(0.125)	(0.444)				
Population2000 (in logs)	0.087	0.389*	0.014	0.367*				
	(0.094)	(0.227)	(0.089)	(0.210)				
Capitalregion	4.164***	1.558***	2.837***	2.988***				
	(0.544)	(0.431)	(0.459)	(1.023)				
Mines (in logs)	0.117*	0.186*	0.003	0.224				
	(0.067)	(0.106)	(0.041)	(0.179)				
Oilgas	0.070	0.326	0.077	0.036				
	(0.149)	(0.438)	(0.133)	(0.807)				
Area (in logs)	0.234**	0.367	0.183**	−0.420				
	(0.091)	(0.233)	(0.080)	(0.497)				
Ports	−0.068	−0.256	−0.155	−1.797*				
	(0.193)	(0.684)	(0.150)	(1.044)				
Roaddensity	1.145	1.406	1.181	4.137*				
	(1.198)	(2.166)	(1.080)	(2.360)				
Country-year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ADM1 FE	No	No	No	No	Yes	Yes	Yes	Yes
R-squared	0.40		0.35		0.30		0.28	
Observations	8327	8327	8375	8375	8327	8327	8375	8375
Regions	709	709	709	709	709	709	709	709

Notes: The dependent variable is Chinese total flows (in logs) in columns 1 and 5, Chinese total flows (in levels) in columns 2 and 6, Chinese ODA-like flows (in logs) in columns 3 and 7, and Chinese ODA-like flows (in levels) in columns 4 and 8. Standard errors (in parentheses) clustered at the country level. *** (**, *): significant at the 1% (5%, 10%) level.

Spatial data

DSIER [/dɪˈzaɪər/] — Summer 2024

Julian Hinz

Bielefeld University