INSTITUTE OF DEVELOPING ECONOMIES

IDE Discussion Papers are preliminary materials circulated
to stimulate discussions and critical comments

IDE DISCUSSION PAPER No. 589

Some notes on the spatial representations

Souknilanh Keola and
Kenmei Tsubota*

Abstract

There are conventional methods to calculate the centroid of spatial units and distances among them with using Geographical Information Systems (GIS). The paper points out potential measurement errors of this calculation. By taking Indian district data as an example, systematic errors concealed in such variables are shown. Two comparisons are examined; firstly, we compare the centroid obtained from the spatial units, polygons, and the centre of each city where its district headquarters locates. Secondly, between the centres represented in the above, we calculate the direct distances and road distances obtained from each pair of two districts. From the comparison between the direct distances of centroid of spatial units and the road distances of centre of district headquarters, we show the distribution of errors and list some caveats for the use of conventional variables obtained from GIS.

Keywords: Centrality, road distance, direct distance
JEL classification: C21, D82,

* *Corresponding author*. Research Fellow at IDE-JETRO and Visiting Research Fellow at Vrije Universiteit Amsterdam. Email: kenmei.tsubota@gmail.com

The Institute of Developing Economies (IDE) is a semigovernmental, nonpartisan, nonprofit research institute, founded in 1958. The Institute merged with the Japan External Trade Organization (JETRO) on July 1, 1998. The Institute conducts basic and comprehensive studies on economic and related affairs in all developing countries and regions, including Asia, the Middle East, Africa, Latin America, Oceania, and Eastern Europe.

The views expressed in this publication are those of the author(s). Publication does not imply endorsement by the Institute of Developing Economies of any of the views expressed within.

INSTITUTE OF DEVELOPING ECONOMIES (IDE), JETRO
3-2-2, WAKABA, MIHAMA-KU, CHIBA-SHI
CHIBA 261-8545, JAPAN

Some notes on the spatial representations[+]

Souknilanh Keola[*] and Kenmei Tsubota[**]

Institute of Developing Economies, JETRO

March, 2016

Abstract

There are conventional methods to calculate the centroid of spatial units and distances among them with using Geographical Information Systems (GIS). The paper points out potential measurement errors of this calculation. By taking Indian district data as an example, systematic errors concealed in such variables are shown. Two comparisons are examined; firstly, we compare the centroid obtained from the spatial units, polygons, and the centre of each city where its district headquarters locates. Secondly, between the centres represented in the above, we calculate the direct distances and road distances obtained from each pair of two districts. From the comparison between the direct distances of centroid of spatial units and the road distances of centre of district headquarters, we show the distribution of errors and list some caveats for the use of conventional variables obtained from GIS.

Keywords: Centrality, road distance, direct distance

JEL classification: C21, D82,

[+] We would like to acknowledge Arif Khan and his team for the data entry and Hihi Sai for her excellent research assistance.

[*] Research Fellow at IDE-JETRO.

[**] Corresponding author. Research Fellow at IDE-JETRO and Visiting Research Fellow at Vrije Universiteit Amsterdam. *Email:* kenmei.tsubota@gmail.com

1. Introduction

Spatial units such as counties, cities, and municipalities, are some of the popular observations for empirical studies. Some of the studies with these observations consider the spatial relations of the observations such as market accessibility, proximity to larger markets, airports, or international ports. For such studies, construction of spatial variables is required.

This paper considers two types of spatial representation of geographical units. One is the centre of the observation and the other is the distance among the observations. There are conventional methods to calculate the centroid of spatial units and distances among them with using Geographical Information Systems (GIS). The paper points out potential measurement errors of such calculations.

By taking Indian district data as an example, systematic errors concealed in such variables are shown by the comparison of two spatial representations. One is the centrality and the other is the distance. Firstly, we compare the centroid obtained from the spatial units, polygons, and the centre of each city where its district headquarters locates. Centroid of a spatial unit can represent itself only if the attributes are distributed uniformly. For example, centroid of British India was called Zero milestone of India and is located at Nagpur, Maharashtra[1]. This point is the centre of the territory but, of course, is not the centre of population within the territory. Having the centre of the district headquarter city as the centre of population, we show the difference between these two centres. Secondly, between the centres represented in the above, we calculate the direct distances and road distances obtained from each pair of two districts. From the comparison between the direct distances of centroid of spatial units and the road distances of centre of district headquarters, we show the distribution of errors and list some caveats for the use of conventional variables obtained from GIS.

For the calculations of centroid and direct distance, several programs are offered by each platform such as R, ArcGIS Qgis or other software. With these programs, it is straightforward to obtain such variables with one or some lines of command. However, it is not guaranteed that such variables contain certain measurement errors stemming out of the assumptions on the representation of

[1] Coordinate is 21.149840 N and 79.080580 E.

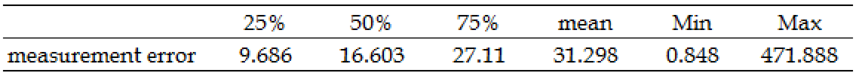
spatial units.

The reminder of the paper is organized as follows. In section 2, we examine the distance between the centroids and the centres of district headquarters as the errors of spatial representation. Section 3 gives the comparison between direct distances of centroids and road distances of centres of district headquarters. The comparison of the distances shows the distribution of errors. Discussion and conclusion appear in section 4.

2. Measurement errors from the centroids

This section shows the systematic bias stemming out from the use of centroid of polygons. When the precise representative locations of administrative boundary are not available or costly, centralities of geographical units are frequently employed as the second best. Centroid of geographical units represents the unweighted centre of it. The use of this centroid implicitly assumes that the variable of interest is uniformly distributed. Thus, if the target variable is not uniformly distributed, the measurement errors are always associated with this representation.

With using Indian districts data, this section shows how such errors are systematically distributed (c.f. such errors are relatively larger when the districts have larger size). There are 592 districts. All of the centres are obtained from India Place Finder[2], which is approximately the centre of highly populated areas. After obtaining the centroid, we calculate the Vincenty-style calculation of distance between the true centre and the centroid of districts. The summary statistics is shown in Table 1. Graphical representation of this measurement error is shown in Figure 1. Taking the Vincenty-style distance as vertical axis and the area size as horizontal axis, scattered plots of shows positive correlations of these variables.



	25%	50%	75%	mean	Min	Max
measurement error	9.686	16.603	27.11	31.298	0.848	471.888

Table 1. Measurement error as the distance between centroid and district centre

[2] http://india.csis.u-tokyo.ac.jp/

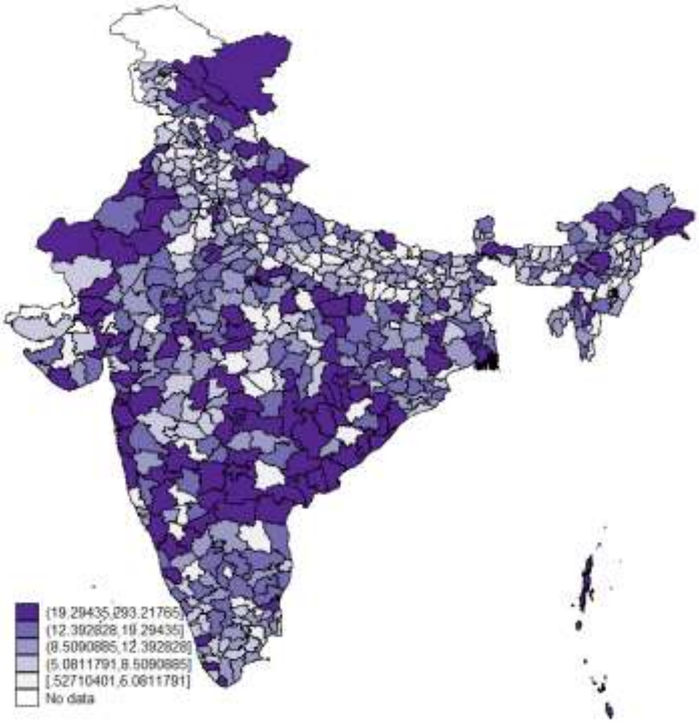


Figure 1. Distance between centroid of polygons and actual centre

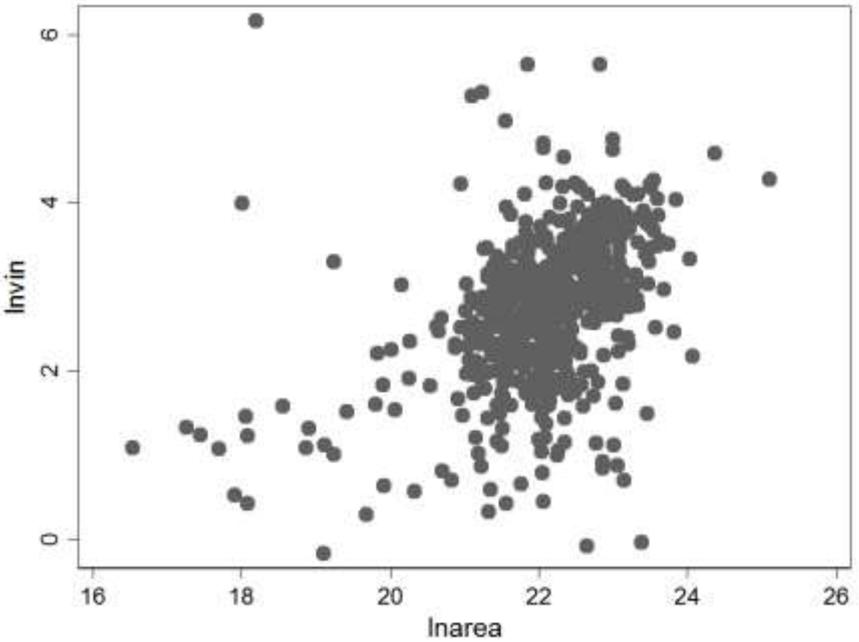
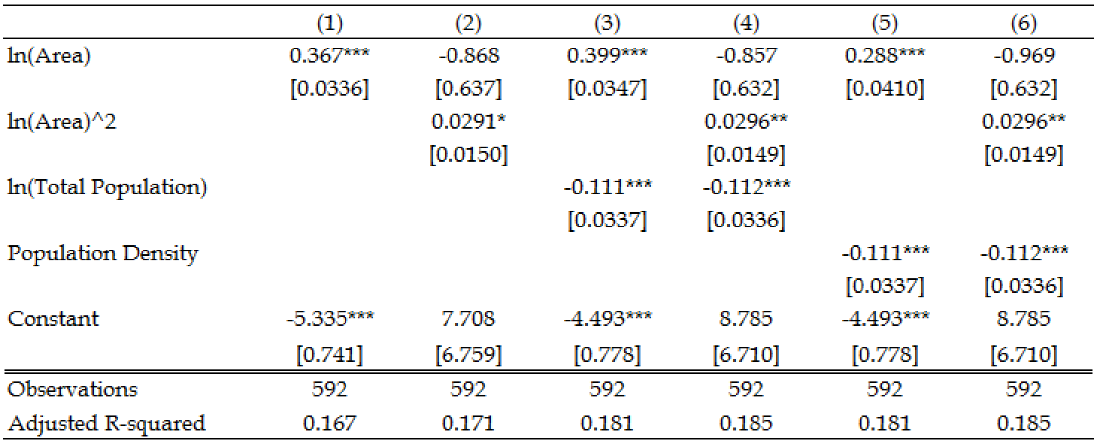


Figure 2. Distance between centroid of polygons and actual centre and its area size

Taking logarithm of area size and population size, we estimate simple OLS regression with some variables. Area size is positive significant. Square of area size is also positive significant when we include it. Inclusion of total population shows negative significant and enlarge the coefficient of area size. However, this significance of total population suggests that the area size is correlated with population size. The inclusion of population density is also in the same direction.



	(1)	(2)	(3)	(4)	(5)	(6)
ln(Area)	0.367***	-0.868	0.399***	-0.857	0.288***	-0.969
	[0.0336]	[0.637]	[0.0347]	[0.632]	[0.0410]	[0.632]
ln(Area)^2		0.0291*		0.0296**		0.0296**
		[0.0150]		[0.0149]		[0.0149]
ln(Total Population)			-0.111***	-0.112***		
			[0.0337]	[0.0336]		
Population Density					-0.111***	-0.112***
					[0.0337]	[0.0336]
Constant	-5.335***	7.708	-4.493***	8.785	-4.493***	8.785
	[0.741]	[6.759]	[0.778]	[6.710]	[0.778]	[6.710]
Observations	592	592	592	592	592	592
Adjusted R-squared	0.167	0.171	0.181	0.185	0.181	0.185

Table 2. Estimation Results: log-log

3. Measurement errors from the calculation of distances

Our arguments have started from the given spatial units, polygons. However, there is a literature of Voronoi diagrams where the locations of observations are given but the boundaries of each unit are not available. In such cases, since the true location may be already given, the errors in the previous section may be negligible. However, there are still worries that the choice of distance calculation may include measurement errors.

In this section, we compare two types of distance; road distance between district HQ and direct distance between centrality of polygon. This is an attempt to compare the measurement errors when the analyst does not have true information. Suppose there is no information on the centrality and road network information, the centrality of polygons and direct distance may be used for the calculation. These constructions are the easiest and always available for scholars. On the other hand, if there is location information as district HQ and road network information, such information and associated road distance may be taken as the variable.

Figure 3 shows the relations between the two distances. Horizontal axis is the road distance and vertical axis is the direct distance (both variables are log).

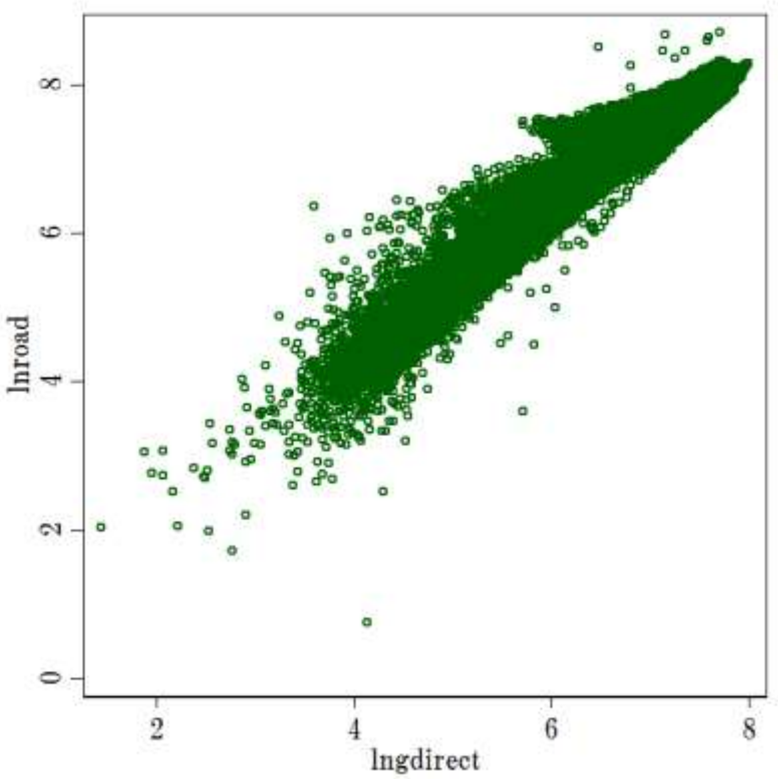
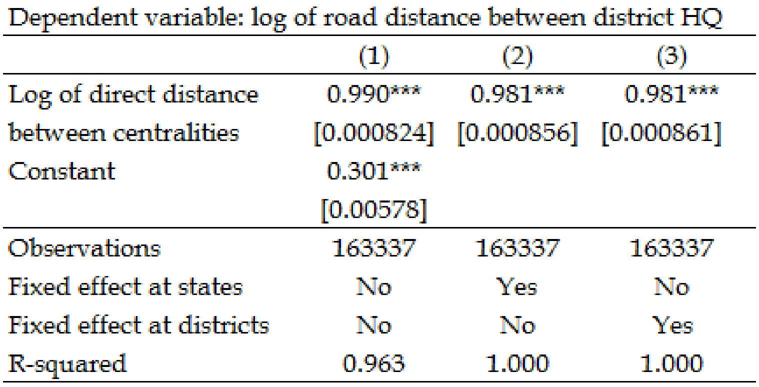


Figure 3. Road distance between district HQ and direct distance between centrality of polygon among Indian districts

As is clearly seen from the Figure 3, both distances are not identical. They are similar. Most of the samples show that road distance is larger than the direct distance. However there are fractions of observations, 4.8%, that direct distance is shorter than road distance. If we compare these two distances at the same location, there is no possibility of having such cases. However, since we use different location information for each distance, this is the source of this result. This may happen when the difference between centrality and the location of HQ is large in origin and destination.



Dependent variable: log of road distance between district HQ			
	(1)	(2)	(3)
Log of direct distance	0.990***	0.981***	0.981***
between centralities	[0.000824]	[0.000856]	[0.000861]
Constant	0.301***		
	[0.00578]		
Observations	163337	163337	163337
Fixed effect at states	No	Yes	No
Fixed effect at districts	No	No	Yes
R-squared	0.963	1.000	1.000

Table 3. Regression of road distance and direct distance

Table 3 shows the regression result. The column 1 does not include fixed effects but column 2 includes fixed effect at states and column 3 includes fixed effect at districts. It shows very high correlation as 0.99 in column 1 but R square is 0.963. This difference is the source of measurement error. When we include state dummy, R square is increased to 1.000 but the coefficient becomes 0.981. It is the same for column 3.

4. Discussion and conclusions

This paper put two notes on the representation of the spatial units with the use of geographical tools. We have examined the possibility of systematic measurement errors by the use of centroid of geographical units. There are two types of possible measurement errors. One is the accuracy of location that appears as the difference between centrality of polygon and the district HQ.[3] The other is the measurement in distance. Typically, direct distance is an easy calculation method of distance. However, it doesn't reflect real road network nor road distance. Road distance is available from Google Map or other web services when the number of observations is not large. If road shapefile is available, it is also possible to obtain road distance.

We have compared two possible measurement errors and found there are always some gaps. Such gaps are not large and are as much as less than 5%.

[3] Throughout this paper, we assumed that the true centre of the district is to be at the location of district HQ. However, if the "true centre" can be defined in the other way, one can find different gaps with it.

However, without district dummy or state dummy to control such variation, any estimation using direct distance suffers from these measurement errors. For further analysis, it may useful to analyze the impact of this measurement error in gravity models of regional trades or other field of studies which heavily use *distance*.