Spatially Explicit Models of City Population Growth in India

Donghwan Kim and Mark R. Montgomery *

Abstract: May 3, 2012

Abstract

This paper presents state-of-the-art econometric models of city population growth in India. Although forecasts of city growth are urgently needed in poor countries to address the challenges they will face in development, environment, and climate change, city-level population forecasting in these countries is still uncommon. However, the data needed to support forecasting are becoming more readily available across a range of poor countries, and in some countries a solid empirical foundation for modelling now exists. India offers not only longitudinal data at the city level for a great number of cities from 1901 to 2001, but also provides disaggregated time-series of important demographic determinants of this growth, state-level urban total fertility rates, child mortality rates, and rates of migration. In methodological terms, this paper combines panel data spatial econometrics with multi-level modelling. We depart from conventional models of spatial correlation in allowing spatial effects to emerge not only from geographic distance, but also from the political and economic context in which the city is situated, with important influences at the state and regional level in India as well as spill-over effects from neighboring cities. To estimate these effects in India, both classical and Bayesian methods are applied. Our initial analysis shows that urban fertility rates display very strong positive effects on city population growth rates and confirms that city growth is spatially correlated. These results will be extended with more additional data and models in which spatial correlation stems from an explicitly multi-level structure.

*Donghwan Kim, Postdoctoral Associate, School of Forestry & Environmental Studies, Yale University, New Haven, donghwan.kim@yale.edu. Mark R. Montgomery, Stony Brook University and Population Council, New York, mmontgomery@popcouncil.org.

1 Overview

This paper presents state-of-the-art econometric models of city population growth in India. Although forecasts of city growth are urgently needed in poor countries to address the challenges they will face in development, environment, and climate change, city-level population forecasting in these countries is still uncommon. Part of the difficulty is that spatially specific data on the determinants of city growth are often lacking. For instance, city-level fertility rates are rarely available in developing countries (including India) although reasonable proxies for the city-level demographic rates are provided by the rates of higher-level administrative units (e.g., states).

This paper develops a method that combines panel data spatial econometrics (Kapoor et al. 2007; Kim 2011) and multi-level modeling (Goldstein 1999). We depart from conventional models of spatial correlation in allowing spatial effects to emerge not only from geographic distance, but also from the political and economic context in which the city is situated, with important influences at the state and regional level in India as well as spill-over effects from neighboring cities.

Efforts to incorporate spatial econometrics into multi-level modeling are still in the formative stage.[1] Corrado and Fingleton (2011) gives a formal discussion of spatial econometrics in multi-level modeling for cross-sectional data, as does Yamagata et al. (2011). To the best of our knowledge, no published paper deals with panel-data spatial econometric models having a multi-level structure.

We apply our new methods to Indian city growth from 1901 to 2001 using an unusually detailed dataset that supplies information on city population growth and its demographic determinants in a spatially explicit manner. Both classical and Bayesian methods are used in estimating the city growth model. Our initial results indicate that urban fertility rates display very strong positive effects on city growth rates in India, and we find that city growth is spatially correlated (Kim 2011). This paper extends the initial analysis with updated data and a multi-level specification.

2 Econometric specifications

To analyze and forecast the growth of India's cities, we first translate each city's series of population counts into a series of growth rates—this can be done for cities with three or more population records—and then link to these growth rates information on urban total fertility rates, child mortality rates, and rates of migration.

[1] In the spatial statistics literature, some papers discuss incorporating spatial effect in multi-level modeling, including Langford et al. (1999), Chaix et al. (2005), Gelfand et al. (2007), and Chaix (2010).

The conversion of the dependent variable from population counts to growth rates may be understood as follows.

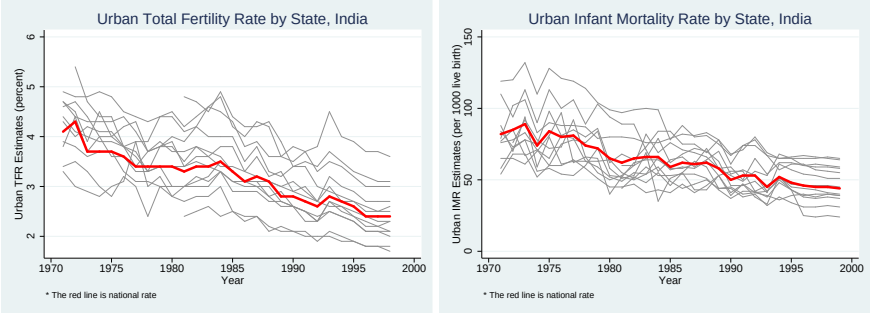
Consider an idealized city with index i, nested in state j, whose boundaries are fixed from time t to $t+1$, and for which $P_{i,j,t}$ and $P_{i,j,t+1}$ are the populations at these two time points. Let $B_{i,j,t}$ represent total births to city i residents in state j from t to $t+1$ and let $D_{i,j,t}$ represent total deaths. Of all those who reside in city i of state j at period t, a total of $O_{i,j,t}$ migrate away from i to all other areas $k \neq i$, and $M_{i,j,t}$ residents of other areas in-migrate to the city. Hence, we can express population growth rate of city i in state j, $g_{i,j,t}$, as

$$g_{i,j,t} \equiv \frac{P_{i,j,t+1} - P_{i,j,t}}{P_{i,j,t}} = b_{i,j,t} - d_{i,j,t} + m_{i,j,t} - o_{i,j,t}, \tag{1}$$

with $b_{i,j,t} = B_{i,j,t}/P_{i,j,t}$, a fertility measure that is not unlike a crude birth rate for the city, and likewise for the measures of mortality, in-migration, and out-migration. Demographers are well aware that $b_{i,j,t}$ is heavily influenced by city i's age and sex composition, as is the $d_{i,j,t}$ mortality measure. Migration rates are also strongly age-dependent and in some contexts also vary importantly by sex. Note that $m_{i,j,t}$ and $o_{i,j,t}$ will tend to vary over i in that the migration sending and receiving areas k that are linked to city i by migration networks—these may be other cities or towns as well as a multitude of rural areas—will differ from one i index to the next.

Reality departs in many ways from this idealized accounting scheme for city population growth, most importantly in that the set of places that are defined to constitute city i can and often will change over the period from t to $t+1$, so that the measured version of $g_{i,j,t}$ will include another component reflecting net population increments over the period (or decrements) produced by boundary changes. Moreover, although the equation above is an accounting identity, its right-hand-side elements are not well measured in readily accessible demographic datasets.

Indeed, nationally representative demographic surveys do not generally supply estimates of demographic rates that are meaningful at the level of individual cities, with capital cities sometimes being an exception. These surveys have also tended to give short shrift to migration, and in particular do not often collect information on the location (by name) of an area from which an in-migrating respondent arrived. Although it is possible to measure urban crude birth rates by combining age data from survey household listings with births data collected from women of reproductive age, reliable measures of crude death rates cannot be collected in this way (information on adult mortality is required) and are generally quite difficult to obtain. In short, a considerable gap separates the fertility, mortality, and migration measures in the accounting scheme of equation (1) from their closest counterparts in accessible empirical data.



(a) Urban total fertility rates (b) Urban infant mortality rates

Figure 1: India urban demographic rates by state along with its national counterparts (in red), 1971 - 1999. Data source: National Commission on Population in India.

Basic Framework

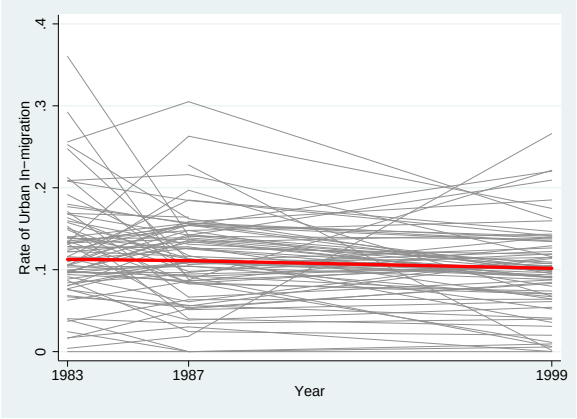
For each city, we have converted population counts for each period t_0 to t_1 given in its data series into continuous growth rates, $g_{i,j,t_0} = (\ln P_{i,j,t_1} - \ln P_{i,j,t_0})/(t_1 - t_0)$. The econometric growth model, inspired by but certainly not identical to the accounting framework spelled out above, is set out as equation (2),

$$g_{i,j,t} = \beta_0 + \beta_1 \mathrm{TFR}_{j,t} + \beta_2 \mathrm{Q}_{j,t} + \mathbf{D}'_{i,j,t}\gamma + \mathbf{X}'_{i,j,t}\delta + v_{i,j,t}. \quad (2)$$

In this equation $g_{i,j,t}$ is the estimated population growth rate for city i in state j at time t, and the fertility and mortality components of growth are represented by the state-level urban total fertility rate $\mathrm{TFR}_{j,t}$ and $Q_{j,t}$, the state-level urban child mortality rate.

In India, as is the case in most developing countries, no city-level demographic rates are available. In their place we employ state-level urban fertility and mortality rates, which for India are available for an impressive span of time. Figure 1(a) shows substantial state-level differences in Indian urban total fertility rates, and also displays the changes in these rates over time[2]. For the illustration presented here, urban total fertility rates serve as our main explanatory variable along with urban infant mortality rates. Our initial analysis which uses national-estimate urban fertility rate shows that urban total fertility rates (TFR) display very strong positive effects on city growth rates (Montgomery 2008; Kim and Montgomery 2011).

[2]We continue our efforts to collect India's disaggregated (i.e. subnational) demographic data including migration. Figure 2(a) shows urban in-migration rates of 80 sub-states for three time periods which we recently collected from IPUMS. Though national-level rate of urban in-migration (in red in the table) is stable over time, regional variation is substantial. This migration component will be incorporated in the future.



(a) Urban in-migration rate

Figure 2: India in-migration rate for 80 sub-states along with it national level (in red), 1983, 1987, and 1999. Data source: IPUMS.

We include in $\mathbf{X}_{i,j,t}$ a set of dummy variables recording city i's population size, which as we will see, turns out to be an important influence on the rate of population growth. We also include here a set of ecosystem indicators, which we discuss in more detail in the next section. The vector $\mathbf{D}_{i,j,t}$ contains a set of dummy variables indicating the start-of-period and end-of-period units in which the city's population is recorded. As will be seen in the next section, city populations recorded in the UN data are defined as different boundary concepts. In principle, of course, a number of additional city-specific explanatory variables could be introduced to explain city growth. Variables that are fixed over time present no particular difficulties; those that change with time, however, would themselves need to be forecast in the process of generating city growth forecasts.

We also need to address the properties of $v_{i,j,t}$, the regression disturbance term. An error-components specification provides a sensible entry-point for our analysis. In such specifications, the disturbance term is represented as a composite,

$$v_{i,j,t} = u_{i,j} + \eta_{i,j,t} \tag{3}$$

containing one component, $u_{i,j}$, that is specific to city i and whose value can be estimated as $\hat{u}_{i,j}$. The estimate $\hat{u}_{i,j}$ is necessary to forecast city growth. To estimate the $u_{i,j}$, Bayesian and classical take different procedure. In Bayesian, $u_{i,j}$ is estimated inside Bayesian Markov Chain Monte Carlo (MCMC) estimation algorithms by considering $u_{i,j}$ as an unknown model parameter (Chib 1996; Kim 2011). In classical approach, the Goldberger (1962)'s best linear unbiased prediction (BLUP) procedure shows how to estimate it (Taub 1979; Baltagi and Li 2004).

Spatial econometrics in multi-level structure

We need to place additional structure on the error terms to account for multiple spatial effects. Equation (4) outlines a specification containing unobservable state effects μ_j which is both time-invariant and city-invariant. For any given state, this term induces correlation in the error terms of the cities of that state, and gives us a means of estimating within- and between-state variances in city growth.

$$v_{i,j,t} = u_{i,j} + \mu_j + \eta_{i,j,t} \tag{4}$$

Equation (5) presents the more conventional Cliff and Ord (1969) type of spatial dependence. In this specification, the disturbance $v_{i,j,t}$ for city i is directly linked, via $\rho w_{i,k}$, to $v_{k,j,t}$, its counterpart for city k. The spatial autocorrelation coefficient ρ and a pre-specified spatial weight $w_{i,k}$ determines the size and direction of the relationship. This model specification has difficulty in estimation when there are missing observations in panel data, that is, panel data are unbalanced (Baltagi et al. 2007; Kim and Montgomery 2011). We use our Fortran programs to estimate the model.

$$v_{i,j,t} = \rho \sum_{k \neq i} w_{i,k} v_{k,j,t} + u_{i,j} + \eta_{i,j,t} \tag{5}$$

Equation (6) is a combined model of equations (4) and (5).

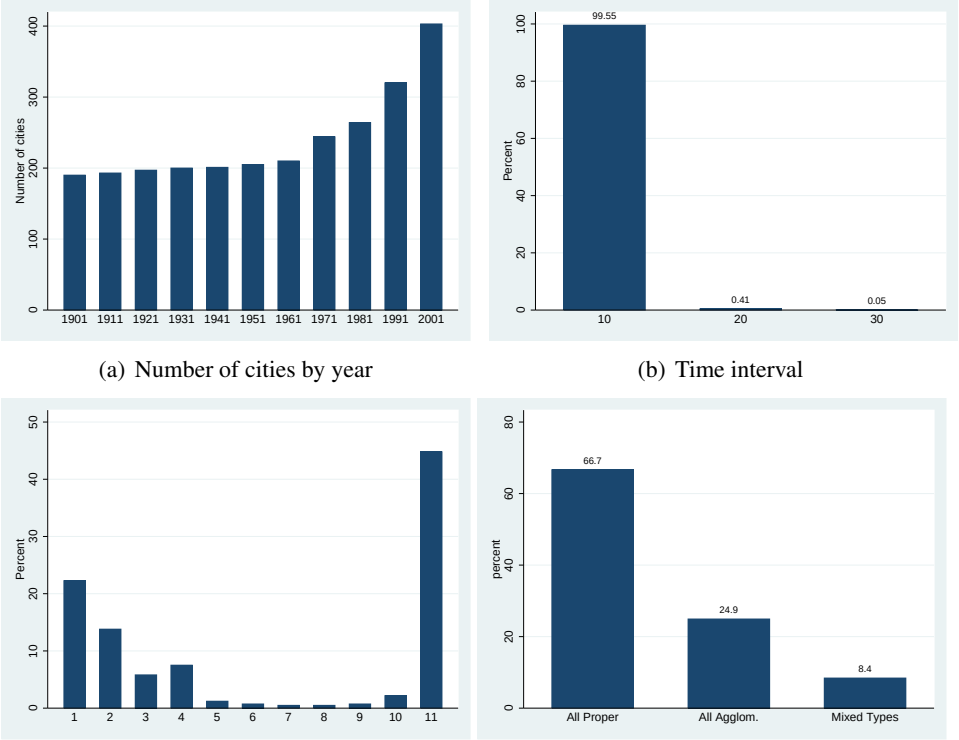
$$v_{i,j,t} = \rho \sum_{k \neq i} w_{i,k} v_{k,j,t} + u_{i,j} + \mu_j + \eta_{i,j,t} \tag{6}$$

3 Data

City Population Data

City population data come from the latest version of the United Nations cities database (United Nations 2010) which is a panel dataset, containing city population counts for thousands of individual cities over time. Though the UN cities database covers cities in almost all of the countries, we limit the scope of our analysis in this paper to a single country, India[3]. For India, the database contains population counts

[3]The UN monitors all cities with populations of 100,000 and above; when a given city crosses this threshold, the Population Division endeavors to reconstruct its history (Montgomery and Balk 2011). The United Nations Population Division is continually expanding and correcting its city population series, devoting substantial effort to the task every two years to prepare the next edition of *World Urbanization Prospects*, and making steady incremental progress in the interim. See Montgomery and Balk (2011) for more details.



(a) Number of cities by year

(b) Time interval

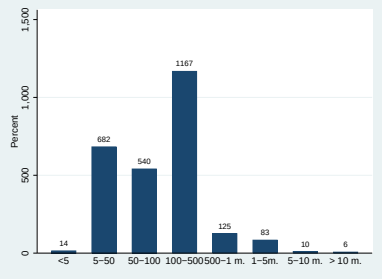
(c) Number of records per city

(d) Statistical concepts on a per-city basis

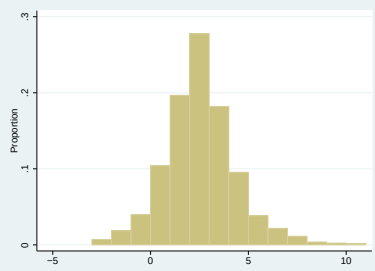
Figure 3: Population records on some 400 cities in India, 1901 - 2001. Data source: UN cities database 2010 version (United Nations 2010).

from 1901 to 2001 for some 400 cities. Figure 3(a) shows the number of cities in the database over time. For any given city, the time interval between records is 10 years with a few exceptions (Figure 3(b)). The number of records varies over city, as summarized in Figure 3(c). For India, city's records are expressed in terms of either *city proper* or *urban agglomeration*. As Figure 3(d) shows, for the India cities with two or more entries in the 2009 database, there are cities – 8.4 percent of the cities – whose statistical concept measuring population changes over time. The city proper is the more common concept in India, with the populations of 66.7 percent of India cities being consistently recorded in this way.

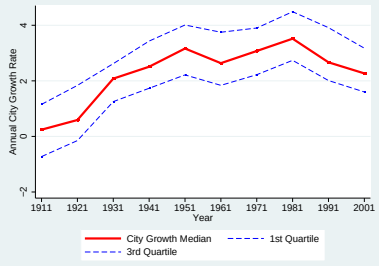
Figure 4(a) depicts the distribution of city population counts in India which are used to calculate growth rates. Of the 2, 627 population counts recorded, about 44 percent fall in the range between 100,000 and 500,000 persons. There are also some records for cities as small as 5,000 in population. Figure 4(c) shows the distribution of city growth rates for all cities and time periods available since 1901. Over this



(a) City population records



(b) City growth rates



(c) City growth rates

Figure 4: City population records and city growth rates, All cities in India, 1901 - 2001, Data source: UN cities database 2010 version (United Nations 2010).

long period, the median growth rate recorded in the dataset is 2.45 percent and the mean is 2.56 percent. As the figure shows, there are instances of city population decline in these data as well as cases of rapid growth at rates of 10 percent and above. Figure **??** shows city growth rate by state.

The right-hand side components

Limiting our city growth analysis to a single country (i.e. India) in this analysis, we test more disaggregated urban vital rates rather than their national-level counterparts. Both state-level estimates of urban fertility and child mortality rates come from National Commission on Population in India (`http://populationcommission.nic.in`). As seen in Table 1(a) and 1(b), they are currently available for the time periods from 1971 to 1999. With the state-level urban vital rates in hand, our panel data analysis is restricted to the time periods where the right-hand components of the city growth econometric models are available though city population data covers more extended time periods from 1901 to 2001.

We also include ecozone variables: inland water, LECZ (low elevation coastal zone), and degree of aridity along with city's average slope and elevation. The ecozone data are in a raster format of geospatial data, which are linked to city growth rate using GIS/geospatial analysis (see Kim (2011) for more details).

4 Results

Basic panel data analysis

Table 1 shows basic panel data analysis of city population growth in India for three time periods 1971, 1981, and 1991. Three basic model specifications are used here: random-effects model, fixed-effects model, and pooled OLS model which have different assumption on the variance-covariance matrix of the regression error terms. Annual percentage growth rate of city population is used as the dependent variable.

As shown in the table, the start-period state urban TFR has very strong positive effects on city growth. The coefficient ranges 0.529 – 1.019 (depending on model specification), meaning that a drop of 1 child in the state-level urban TFR is associated with a drop of 0.529 to 1.019 percentage point in the rate of city growth in India. The fixed-effects estimate of the total fertility rate coefficient is by far the largest in this set of estimates – This result is coincident with our previous result. The start-period urban child mortality also has significant effect on city growth though the effect is small.

The results show that city size is also an important determinant of city growth. Larger cities tend to grow more slowly than do cities under 100,000 population (which is the omitted category in the regression specification). Cities with 100,000 to 500,000 persons experience lower growth rate of 0.694 – 0.725 percentage point than cities under 100,000 persons.

The effects of ecological and geophysical characteristics are not straightforward in India. These are time-invariant dummy variables, so the effects can be estimated with pooled OLS and random-effects specifications. The inland water dummy variable has positive sign but its statistical significance does not hold. However, cities in LECZ experience higher growth rate of 0.591 – 0.621 percentage point than non-LECZ cities. The results show that the degree of aridity, slope, and elevation have no significant impact on city growth in India.

Table 1: Basic city growth regression models, Cities in India. 1971-1991.

	Model 1			Model 2			Model 3		
	OLS	FE	RE	OLS	FE	RE	OLS	FE	RE
State Urban TFR	0.562	1.019	0.665	0.529	0.839	0.600	0.563	0.839	0.622
(Z statistic)	(4.88)	(6.38)	(5.65)	(4.65)	(4.93)	(5.17)	(4.80)	(4.93)	(5.27)
State Urban Q5	-0.007	-0.016	-0.008	-0.009	-0.016	-0.010	-0.008	-0.016	-0.009
	(-1.73)	(-2.45)	(-1.83)	(-2.20)	(-2.50)	(-2.23)	(-1.87)	(-2.50)	(-2.02)
100 <= City Size < 500				-0.723	-0.693	-0.724	-0.725	-0.693	-0.722
				(-4.39)	(-3.00)	(-4.36)	(-4.40)	(-3.00)	(-4.36)
500 <= City Size < 1,000				-0.664	-1.083	-0.744	-0.652	-1.083	-0.728
				(-2.45)	(-2.41)	(-2.63)	(-2.37)	(-2.41)	(-2.56)
City Size >= 1,000				-0.216	-1.242	-0.400	-0.305	-1.242	-0.456
				(-0.68)	(-1.97)	(-1.15)	(-0.95)	(-1.97)	(-1.32)
Inland Water							0.091		0.114
							(0.64)		(0.68)
LECZ							0.591		0.621
							(3.18)		(2.86)
Dry subhumid							-0.111		-0.131
							(-0.69)		(-0.70)
Semiarid							0.290		0.272
							(1.71)		(1.38)
Arid and above							-0.223		-0.250
							(-0.29)		(-0.29)
Slope							0.003		0.002
							(1.70)		(1.25)
Elevation							0.001		0.001
							(1.45)		(1.24)
Constant	1.460	0.744	1.165	2.295	1.859	2.133	1.682	1.859	1.580
	(5.38)	(1.03)	(4.09)	(6.34)	(2.27)	(5.51)	(4.09)	(2.27)	(3.59)
σ_u			0.792			0.763			0.716
			(9.75)			(9.46)			(8.77)
σ_η			1.299			1.284			1.284
			(28.19)			(28.19)			(28.18)

The number of observations is 648. The dummy variables for changes in city definition are not included in the table. The baseline category with city defined in terms of urban agglomeration at start and end of spell.

Analysis of spatial effects: Multilevel modeling

Table 2 shows Bayesian results of city growth panel data models with multi-level structure. We use a 2-level structure with city and state. Spatial effect stemming from multilevel nested structure is seldom considered in panel data econometric analysis. In India, district is more lower-level administrative unit than state. However, most of districts contain only one city in the data, the district level is not considered.

The model combines panel data econometric technique and multilevel modeling technique each of which has been developed separately – (the former in economics and the latter in social science, especially in education and geographic studies). For Bayesian, both the fixed-effects and random-effects specifications for the 2 levels (i.e. city and state) are used. In multileveling literature, the fixed-effects specification, here for state-specific effects, is seldom used. Table 3 shows their classical results (only for random-effects specifications).

Analysis of spatial effects: Spatial econometrics

Table 4 shows Bayesian results of city growth spatial econometric models in India. In methodological term, it is a panel data random-effects model with spatial correlated errors when panel data is unbalanced. The unbalancedness of panel data gives additional technical difficulty in estimating in this model specification since the matrix of spatial weights, $w_{i,k}$ in Equation (5), differs over time. This model is developed by Kim and Montgomery (2011) for Bayesian and Baltagi et al. (2007) for classical.

The model needs geographic information (e.g. latitude and longitude coordinates) of cities to specify the spatial weights $w_{i,k}$ between cities i and k. Let $d_{i,k}$ be a distance between city centroids. We use spatial weights specified as row-standardized version of inverse distance, $w_{i,k} = d_{i,k}^{-\alpha} / \sum_{k=1}^{N_t} (d_{i,k}^{-\alpha})$ with a parameter α where N_t is the number of city observations at time t. This specification implies that the linkage between the growth rate disturbance terms of cities i and k grows weaker the more distance the two cities are. Also, the degree of weakness of linkage depends on the parameter α ($\alpha > 0$). Distances are expressed in kilometers, measured by formula of the Haversine great-circle distance.

Table 4 shows results when we set α as either 2 or 1. When $\alpha = 2$, the spatial autoregressive coefficient ρ is positive, 0.493 or 0.482, indicating that city growth disturbance is positively correlated. The start-period urban TFR is still statistically significant but the coefficient becomes smaller than one when $\rho = 0$. We can this happening in other variables in the models. When When $\alpha = 1$, ρ is very high, 0.903 or 0.899, and urban TFR has very small values of coefficient and lose statistical significance. However, other variables have similar results as those when $\alpha = 2$.

Table 2: Bayesian city growth panel data models with 2-level (city-state) structure, Cities in India. 1971–1991

	Model 2		Model 3
	2-level FE	2-level RE	2-level RE
State Urban TFR	0.845	0.7223	0.7067
	(0.1764)	(0.1504)	(0.1481)
State Urban Q5	-0.01636	-0.01104	-0.01089
	(0.00649)	(0.005745)	(0.005701)
100 <= CitySize < 500	-0.6888	-0.608	-0.6447
	(0.233)	(0.1643)	(0.1662)
500 <= CitySize < 1,000	-1.047	-0.5477	-0.6184
	(0.4491)	(0.2757)	(0.2845)
CitySize >= 1,000	-1.203	-0.2467	-0.3624
	(0.6304)	(0.3309)	(0.3434)
Inland Water			0.2151
			(0.1665)
LECZ			0.2562
			(0.2344)
Dry subhumid			-0.04054
			(0.2216)
Semiarid			0.258
			(0.2371)
Arid and above			0.05151
			(0.8877)
Slope			0.002271
			(0.001874)
Elevation			3.164E-5
			(5.355E-4)
Constant		1.754	1.468
		(0.4698)	(0.5212)
Standard deviation of state-specific effects σ_μ		0.5424	0.501
		(0.1548)	(0.1539)
Standard deviation of city-specific effects σ_u		0.6156	0.6242
		(0.09072)	(0.09313)
σ_η	1.278	1.294	1.295
	(0.04519)	(0.0465)	(0.04671)
DIC (deviance information criterion)	2401.53	2277.12	2282.62

Posterior mean and posterior standard deviation (in parentheses) are shown in the table.

Table 3: Classical city growth random-effects models with 2-level (i.e. city-state) structure, Cities in India. 1971–1991

	Model 1	Model 2	Model 3
State Urban TFR	0.823	0.722	0.709
	(5.73)	(5.02)	(4.99)
State Urban Q5	-0.011	-0.011	-0.011
	(-1.87)	(-1.96)	(-1.95)
100 <= CitySize < 500		-0.610	-0.645
		(-3.73)	(-3.90)
500 <= CitySize < 1,000		-0.552	-0.621
		(-2.00)	(-2.20)
CitySize >= 1,000		-0.254	-0.365
		(-0.77)	(-1.07)
InlandWater			0.219
			(1.32)
LECZ			0.249
			(1.09)
Dry subhumid			-0.046
			(-0.21)
Semiarid			0.248
			(1.04)
Arid and above			0.051
			(0.06)
Slope			0.002
			(1.20)
Elevation			0.000
			(0.05)
Constant	0.883	1.752	1.466
	(2.50)	(3.84)	(2.87)
Standard deviation of state-specific effects σ_μ	0.565	0.516	0.482
	(3.88)	(3.78)	(3.46)
Standard deviation of city-specific effects σ_u	0.649	0.633	0.641
	(7.72)	(7.46)	(7.45)
σ_η	1.292	1.286	1.286
	(28.38)	(28.29)	(28.26)

Table 4: Bayesian city growth regression models with spatially correlated errors, Cities in India. 1971 - 1991. Spatial weights are $1/d_{i,k}^{\alpha}$ where $d_{i,k}$ is the Haversine great-circle distance between cities i and k. Spatial weights are row-standardized.

	Model 1		Model 2	
	$\alpha = 2$	$\alpha = 1$	$\alpha = 2$	$\alpha = 1$
State Urban TFR	0.568	0.063	0.525	0.062
	(3.47)	(0.35)	(3.31)	(0.35)
State Urban Q5	-0.011	-0.012	-0.012	-0.012
	(-2.05)	(-2.30)	(-2.21)	(-2.46)
100 <= CitySize < 500			-0.574	-0.525
			(-3.65)	(-3.22)
500 <= CitySize < 1,000			-0.604	-0.490
			(-2.24)	(-1.80)
CitySize >= 1,000			-0.182	-0.060
			(-0.54)	(-0.18)
Constant	1.79	3.406	2.458	3.908
	(3.70)	(2.59)	(4.77)	(3.05)
Spatial autoregressive coefficient ρ	0.493	0.903	0.482	0.899
	(7.68)	(18.74)	(7.21)	(18.37)
Standard deviation of city-specific effects σ_u	0.765	0.771	0.750	0.749
	(10.11)	(10.55)	(9.82)	(9.66)
σ_ε	1.23	1.220	1.227	1.215
	(28.22)	(29.17)	(27.90)	(27.76)

Note: Results come from Bayesian MCMC samples. Bayesian Z-statistics in parentheses which is calculated by dividing posterior mean by posterior standard error. Controls for city definition included, but the coefficients are not shown.

References

Badi H Baltagi and Dong Li. Prediction in the panel data model with spatial correlation. In Raymond J.G.M. Florax Luc Anselin and Sergio J. Rey, editors, *Advances in Spatial Econometrics: Methodology, Tools and Applications*, pages 283–295. Springer-Verlag, 2004.

Badi H. Baltagi, Peter Egger, and Michael Pfaffermayr. Estimating models of complex FDI: Are there third-country effects? *Journal of Econometrics*, 140: 260–281, 2007.

Basile Chaix. Modeling effects of the social and physical environment on health: a spatial perspective. Presentation at EPA (United States Environmental Protection Agency), Available at http://www.epa.gov/ncer/events/calendar/2010/mar17/presentations/chaixe.pdf, 2010.

Basile Chaix, Juan Merlo, and Pierre Chauvin. Comparison of a spatial approach with the multilevel approach for investigating place effects on health: the example of healthcare utilisation in france. *Journal of Epidemiology & Community Health*, 59:517–526, 2005.

Siddhartha Chib. Inference in panel data models via gibbs sampling. In Lszl Mtys and Patrick Sevestre, editors, *The Econometrics of Panel Data: A Handbook of the Theory with Applications Series*. Kluwer Academic Publishers, 1996.

Andrew David Cliff and J. Keith Ord. The problem of spatial autocorrelation. In A. J. Scott, editor, *London Papers in Regional Science, Studies in Regional Science*, page pp. 2555. Pion, London, 1969.

Luisa Corrado and Bernard Fingleton. Multilevel modelling with spatial effects. Working Paper of University of Strathclyde Business School, February 2011.

Alan E. Gelfand, Sudipto Banerjee, C.F. Sirmans, Yong Tu, and Seow Eng Ong. Multilevel modeling using spatial processes: Application to the singapore housing market. *Computational Statistics & Data Analysis*, 51:3567–3579, 2007.

Arthur S. Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57:369–375, 1962.

Harvey Goldstein. *Multilevel Statistical Models*. Wiley, 1999. Available at www.soziologie.uni-halle.de/langer/multilevel/books/goldstein.pdf.

Mudit Kapoor, Harry H. Kelejian, and Ingmar R. Prucha. Panel data models with spatially correlated error components. *Journal of Econometrics*, 140:97–130, 2007.

Donghwan Kim. *Econometric Modeling of City Population Growth in Develping Countries*. PhD thesis, Department of Economics, State University of New York at Stony Brook, 2011.

Donghwan Kim and Mark R. Montgomery. An econometric approach to forecasting city population growth in developing countries. Working paper, Department of Economics, Stony Brook University, 2011.

Ian H. Langford, Alastair H. Leyland, Jon Rasbash, and Harvey Goldstein. Multilevel modeling of the geographic distributions of diseases. *Journal of Royal Statistical Society, Series C (Applied Statistics)*, 48:253–268, 1999.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Mark R. Montgomery. The urban transformation of the developing world. *Science*, 319:761–764, 2008.

Mark R. Montgomery and Deborah Balk. The urban transition in developing countries: Demography meets geography. In E. Birch and S. Wachter, editors, *Global Urbanization*. University of Pennsylvania Press, 2011.

Keith Ord. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126, 1975.

Allan J. Taub. Prediction in the context of the variance-components model. *Journal of Econometrics*, 10:103–107, 1979.

Luke Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

United Nations. *World Urbanization Prospects: The 2009 Revision. CD-ROM Edition - Data in digital form.* United Nations, Population Division, Department of Economics and Social Affairs, 2010.

Yoshiki Yamagata, Hajime Seya, Daisuke Murakami, and Morito Tsutsumi. Hedonic analysis of environmental factors and disaster risk using a multi-level spatial econometric model. Presented in the Vth World Conference of Spatial Econometrics Association, Toulouse, July 6-8, 2011.

A Implementing and developing models

Most of models in the paper are implemented with WINBUGS for Bayesian and STATA for classical except the spatial econometric model. Below shows Bayesian random-effects model with spatial correlation when panel data are unbalanced. It is re-printed here from our previous work (Kim and Montgomery 2011). It is programmed with Fortran 95.

Random-effects model with spatial correlation

Here we describe the Bayesian estimation method for unbalanced panel data with both random effects and spatially correlated disturbances. Consider $\mathbf{g}_t$, a vector of growth rates of all cities available in time t whose dimension N_t is the number of observations for time t, with $t = 1,\ldots,T$. (Note that the data are now ordered differently from what was assumed above.) For each time t the city growth model is written as

$$\begin{aligned}\mathbf{g}_t &= \mathbf{X}_t\theta + \mathbf{v}_t \\ \mathbf{v}_t &= \rho\mathbf{W}_t\mathbf{v}_t + \varepsilon_t \\ \varepsilon_t &= \mathbf{D}_t\mathbf{u} + \eta_t\end{aligned}$$

in which the spatial weight matrix $\mathbf{W}_t$ is of dimension $N_t \times N_t$. Its diagonal elements are all zeros and its off-diagonal elements are w_{ij}. The vector $\mathbf{u} = (u_1,\cdots,u_N)'$ is an $N \times 1$ vector of random effects, and the matrix $\mathbf{D}_t$ is of dimension $N_t \times N$ which is obtained from an $N \times N$ identity matrix by extracting the rows corresponding to cities that provide records at time t (Baltagi et al. 2007). Assume that $\eta_t \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2\mathbf{I}_t)$.

The posterior distribution of the model, $p(\theta, \mathbf{u}, \sigma_u^2, \sigma_\eta^2, \rho|\mathbf{g}, \mathbf{X})$, can be simulated using a Metropolis-within-Gibbs algorithm (Tierney 1994), which is a combination of the Gibbs sampler and Metropolis-Hastings methods. The priors are essentially the same as those of the random-effects model, with the addition of $\sigma_\eta^2 \sim iG(v_0/2, s_0/2)$ and an improper prior for ρ.

Using the block (θ), $\sigma_{\mathbf{u}}^2$, σ_η^2, and ρ, the Metropolis-within-Gibbs algorithm proceeds as follows:

- Define $\mathbf{B} = \mathbf{I}_n - \rho\mathbf{W}_n$ with $\mathbf{W}_n = diag(\mathbf{W}_1,\cdots,\mathbf{W}_T)$, $\widetilde{\mathbf{X}} = \mathbf{B}\mathbf{X}$ with $\mathbf{X} = (\mathbf{X}_1^{'},\cdots,\mathbf{X}_T^{'})'$, and $\widetilde{\mathbf{g}} = \mathbf{B}\mathbf{g}$ with $\mathbf{g} = (\mathbf{g}_1^{'},\cdots,\mathbf{g}_T^{'})'$.
- Define $\mathbf{M}_1 = \mathbf{M}_0 + \sigma_\eta^{-2}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, and $\mathbf{D} = (\mathbf{D}_1^{'},\cdots,\mathbf{D}_t^{'})'$.
- Draw θ from

$$\theta \sim \mathcal{N}\left(\mathbf{M}_1^{-1}\left(\mathbf{M}_0\theta_0 + \sigma_\eta^{-2}\widetilde{\mathbf{X}}'(\widetilde{\mathbf{g}} - \mathbf{D}\mathbf{u})\right), \mathbf{M}_1^{-1}\right).$$

- Define, for each i, $\overline{\widetilde{g}}_i = n_i^{-1}\sum_t \widetilde{g}_{it}$ and $\overline{\widetilde{\mathbf{X}}}_i = n_i^{-1}\sum_t \widetilde{\mathbf{X}}_{it}$.
- Draw u_i from

$$u_i \sim \mathcal{N}\left(\frac{n_i\sigma_u^2}{n_i\sigma_u^2+\sigma_\eta^2}(\overline{\widetilde{g}}_i - \overline{\widetilde{\mathbf{X}}}_i\theta), \frac{\sigma_u^2\sigma_\eta^2}{n_i\sigma_u^2+\sigma_\eta^2}\right).$$

- Draw σ_η^2 from

$$\sigma_\eta^2 \sim iG\left(\frac{n+v_0}{2}, \frac{(\widetilde{\mathbf{g}}-\widetilde{\mathbf{X}}\theta-\mathbf{Du})'(\widetilde{\mathbf{g}}-\widetilde{\mathbf{X}}\theta-\mathbf{Du})+s_0}{2}\right).$$

- Draw σ_u^2 from

$$\sigma_u^2 \sim iG\left(\frac{N+h_0}{2}, \frac{\mathbf{u}'\mathbf{u}+p_0}{2}\right)$$

The kernel of the full conditional posterior distribution of ρ is as follows.

$$p(\rho|\theta,\mathbf{u},\sigma_u^2,\sigma_\eta^2,\mathbf{g},\mathbf{X}) \propto |\mathbf{B}|\exp\left(-\frac{1}{2\sigma_\eta^2}(\widetilde{\mathbf{g}}-\widetilde{\mathbf{X}}\theta-\mathbf{Du})'(\widetilde{\mathbf{g}}-\widetilde{\mathbf{X}}\theta-\mathbf{Du})\right)$$

The random-walk Metropolis-Hastings algorithm for ρ draws a candidate ρ^* from a candidate-generating function, here, a (truncated) normal distribution: at the $i+1$-th iteration, ρ^* is drawn from

$$\rho^* \sim \mathcal{N}(\rho_i, c^2)$$

in which ρ_i is the draw from the previous (i-th) iteration and c is a tuning parameter that is used to adjust the acceptance rate of the MH algorithm. We then calculate the ratio $p(\rho^*)/p(\rho)$ in which $p(\cdot)$ is the kernel of the full conditional for ρ. If the ratio is greater than 1, the candidate is accepted (that is, $\rho_{i+1} = \rho^*$). If the ratio is less than 1, however, the candidate is accepted with probability $p(\rho^*)/p(\rho)$; that is, we take an uniform $(0,1)$ random number u and if $u < p(\rho^*)/p(\rho)$, accept the candidate and if $u > p(\rho^*)/p(\rho)$, do not accept it, in the latter case leaving $\rho_{i+1} = \rho_i$). The set of draws behaves like the draws from the the full conditional posterior distribution of ρ. This was the idea of Metropolis et al. (1953) which revolutionized Bayesian inference.

In practice, we have used the natural logarithm of $p(\cdot)$ which includes the log-determinant, $\ln|B|$. Ord (1975) showed that $|\mathbf{I}-\rho\mathbf{W}_n| = \prod_{i=1}^n(1-\rho\lambda_i)$ with λ_i being the i-th eigenvalue of the spatial weight matrix $\mathbf{W}_n$ of dimension n. It is computationally efficient to use the fact that the eigenvalues of the block-diagonal matrix $\mathbf{W}_n = diag(\mathbf{W}_1,\cdots,\mathbf{W}_T)$ are those of the diagonal blocks $\mathbf{W}_1,\cdots,\mathbf{W}_T$. . In our analysis, we specify vague priors for the hyperparameters; that is, $\theta_0 = \mathbf{0}$, $\mathbf{M}_0 = 1^{-5}\mathbf{I}$, $h_0 = 0$, $p_0 = 0$, $v_0 = 0$, and $s_0 = 0$.

B Results from Initial Analysis

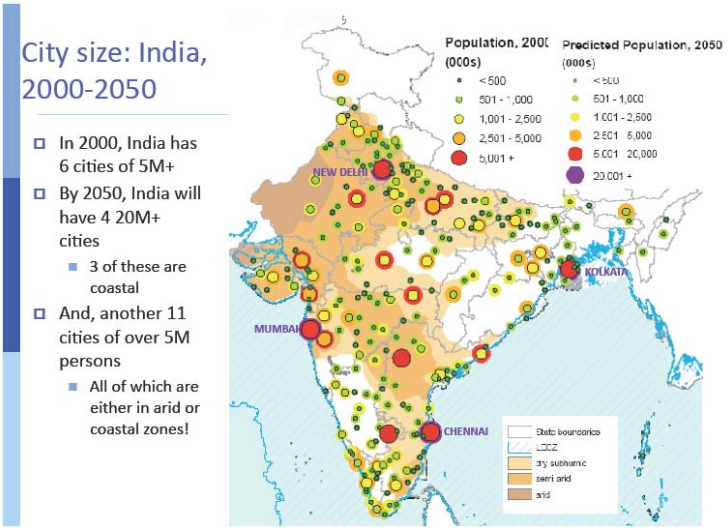


Figure 5: Map of estimated and projected India city size, 2000 and 2050.