City-Size Distribution as a Function of Socioeconomic Conditions: An Eclectic Approach to Downscaling Global Populations

Kyung-Min Nam and John M. Reilly

Report No. 213
March 2012

The MIT Joint Program on the Science and Policy of Global Change is an organization for research, independent policy analysis, and public education in global environmental change. It seeks to provide leadership in understanding scientific, economic, and ecological aspects of this difficult issue, and combining them into policy assessments that serve the needs of ongoing national and international discussions. To this end, the Program brings together an interdisciplinary group from two established research centers at MIT: the Center for Global Change Science (CGCS) and the Center for Energy and Environmental Policy Research (CEEPR). These two centers bridge many key areas of the needed intellectual work, and additional essential areas are covered by other MIT departments, by collaboration with the Ecosystems Center of the Marine Biology Laboratory (MBL) at Woods Hole, and by short- and long-term visitors to the Program. The Program involves sponsorship and active participation by industry, government, and non-profit organizations.

To inform processes of policy development and implementation, climate change research needs to focus on improving the prediction of those variables that are most relevant to economic, social, and environmental effects. In turn, the greenhouse gas and atmospheric aerosol assumptions underlying climate analysis need to be related to the economic, technological, and political forces that drive emissions, and to the results of international agreements and mitigation. Further, assessments of possible societal and ecosystem impacts, and analysis of mitigation strategies, need to be based on realistic evaluation of the uncertainties of climate science.

This report is one of a series intended to communicate research results and improve public understanding of climate issues, thereby contributing to informed debate about the climate issue, the uncertainties, and the economic and social implications of policy alternatives. Titles in the Report Series to date are listed on the inside back cover.

Ronald G. Prinn and John M. Reilly
Program Co-Directors

For more information, please contact the Joint Program Office

Postal Address: Joint Program on the Science and Policy of Global Change
77 Massachusetts Avenue
MIT E19-411
Cambridge MA 02139-4307 (USA)

Location: 400 Main Street, Cambridge
Building E19, Room 411
Massachusetts Institute of Technology

Access: Phone: +1(617) 253-7492
Fax: +1(617) 253-9845
E-mail: globalchange@mit.edu
Web site: http://globalchange.mit.edu/

Printed on recycled paper

City-Size Distribution as a Function of Socioeconomic Conditions: An Eclectic Approach to Downscaling Global Population

Kyung-Min Nam*† and John M. Reilly*

Abstract

In this study, we introduce a new method of downscaling global population distribution, for which purpose conventional approaches have serious limitations in application. Our approach is "eclectic," as it explores the intersection between an optimization framework and the empirical regularities involved in rank-size distributions. The novelty of our downscaling model is that it allows city-size distributions to interact with socioeconomic variables. Our contribution to the urban studies literature is twofold. One is our challenge to the conventional view that the proportionate growth dynamics underlies empirical rank-size regularities. We first show that the city-size distribution of a region can deviate substantially from a log-normal distribution with cross-regional and time variations, and then demonstrate that such variations can be explained by certain socioeconomic conditions that each region confronts at a particular time point. In addition to expanding academic debates on city-size distributions, our study can pave the way for various academic and professional research projects, which need spatial distribution of global population at fine grid cell levels as key input. Our model is applicable to the entire globe, including regions for which reliable sub-regional population data sets are limitedly available, and can be extended easily to function as a forecasting model.

Contents

1. INTRODUCTION

This paper aims to develop a method of estimating global population distribution for a long-term time frame. We are motivated to this study, as traditional approaches based on Zipf's law or economic optimization have significant limits in application to downscaling global population

* Joint Program on the Science and Policy of Global Change, Massachusetts Institute of Technology, Cambridge, MA, USA.

† Corresponding author (Email: kmnam@mit.edu).

distribution at fine grid cell levels, although such detailed information on demographic geography is often required as an essential input for various academic and professional research projects. The existing methods adopt two very different conceptual lenses, where the size of a certain urban agglomeration is either an outcome of a random process embedded in nature or a result of an economy-wide optimization. As will be discussed in detail in Section 2, each of these two frameworks has limitations in application, which arise from the following two opposing reasons.

On the one hand, methods built on Zipf's law attempt to explain spatial distribution of population in isolation from socio-economic dynamics, which may underlie the distribution. The delinking suggests that socio-economic variables do not add much useful information to explain demographic geography within a region. This very fact makes this approach more appropriate for descriptive analysis than for predictive analysis, given that, for example, urbanization, which leads to major changes in the existing pattern of spatial distribution of population, is widely believed to be a function of economic development. Another serious issue involved in this approach is that the robustness of the rank-size relationship depends substantially on the definition of a city, which varies across countries and where researchers have discretion in choosing city limits, for example, metropolitan statistical areas or other definitions of an urban area.

On the other hand, the optimization approach, based on theoretical considerations that would determine an optimal city size, requires data that are not readily available for a comprehensive set of cities, although it is capable of predictive analysis. Any models built on this approach would need extensive data sets for various levels of economies at multiple time points and reliable estimates of many key economic variables and parameters. Constructing such data sets, if assumed feasible, need too much time and energy, and producing reliable estimates for key economic variables is either nearly impossible or a huge task as it is.

Our study attempts to address the issues mentioned above by taking an alternative approach, originally proposed by Asadoorian (2008). Our method is "eclectic," as it is constructed on the conceptual lens that empirical regularities involved in city-size distribution can be expressed as a certain functional form between the size of a city and its rank, but the functional form is determined largely by certain socio-economic variables. In this sense, our major contribution to the urban studies literature is that we demonstrate that a theoretical connection can be built between the two seemingly irreconcilable traditional approaches, and by doing so we introduce a downscaling model that has high applicability and is capable of both descriptive and predictive analyses of spatial distribution of global population.

Our paper is organized in the following way. The first two sections explain the motivation and goal of our research, and review existing literature. Then, the next two sections provide a detailed description of our methodology and empirical analysis results. Section 5 shows how our method can be applied to predictive analysis. Finally, Section 6 synthesizes our conclusions.

2. LITERATURE REVIEW

2.1 City Size Distribution as Random Process

As stated earlier, two approaches have been taken to estimating the distribution of urban population. One of them views urbanization as a random process, which is largely irrelevant to economic dynamics, and focuses on finding a mathematical formulation that fits well to the predetermined outcome of the random process. This approach has been inspired by George Zipf's finding on the quasi-natural regularity underlying urban hierarchies (Zipf, 1949). The so-called Zipf's law or rank-size rule posits that the size of a particular city tends to be determined by the rank of the city in urban hierarchies, sorted in population size, and the size of the prime city. A substantial body of the literature found that the rank-size regularity, proposed by Zipf's law or its modified versions, holds and is robust in many countries over time (Guérin-Pace, 1995; Gabaix, 1999; Petrakos *et al*., 2000; Song and Zhang, 2002; Ioannides and Overman, 2003; Soo, 2007). With such substantial empirical support, Zipf's law has attracted substantial attention in the urban studies field, leading to numerous efforts to explore improved curve fits to city size distribution beyond Zipf's original Pareto distribution function (Giesen *et al*., 2010; González-Val, 2010).

This approach, however, has confronted several challenges. One of them is that the approach built on Zipf's law lacks theoretical foundations (Carroll, 1982; Suarez-Villa, 1988; Fujita *et al*., 1999). Zipf's law is no more than a simple presentation of empirical rank-size regularities, which in itself does not provide any explanation of the main drivers underlying such regularities. Rank-size regularities are often translated into the steady-state outcome of a random process and have been associated with Gibrat's law or the proportionate growth process, which postulates that urban growth is independent of city size (Dobkins and Ioannides, 2000; Gabaix and Ioannides, 2004). While some have found that Gibrat's law holds in reality, others have found conflicting evidence, suggesting that cities with better geographic conditions or market potentials tend to grow faster than others (Rosen and Resnick, 1980; Black and Henderson, 2003; Córdoba, 2008).

Another problem that limits applicability of the approach is that the rank-size rule is not robust enough, particularly when small cities are included in analysis. Zipf's law, positing a log-linear relationship between the size of a city and its rank or a Pareto distribution of city size, tends to hold for the largest cities. In this sense, the definition of a city itself—i.e., the cut-off population level to be considered a city—is a key determinant of the model's applicability (Rosen and Resnick, 1980; Nitsch, 2005). From their cross-sectional study of four regions, for example, Malacarne *et al*. (2001) found that the log-linear relationship does not hold when cities with populations below 100,000 are included. A more complete list of cities tends to have a log-normal distribution in terms of city size, and a Pareto distribution roughly fits to a fraction of the log-normal distribution's upper tail (Eeckhout, 2004). The log-normal distribution of city size can be drawn only when it satisfies Gibrat's law (Gabaix, 1999).

2.2 City Size as Optimization Outcome

In contrast to the approach following in the footsteps of Zipf's law, the other approach looks at a city's size as an outcome of an optimization problem rather than a random process determined by the preexisting urban hierarchy. This approach attempts to capture various interactions among economic actors and variables, often under the assumption that the size of a given city is determined at a level where efficiency of the city is maximized (Fujita *et al*., 1999). Although some early studies seek universal optimality in city size while ignoring various socio-economic conditions that each city may confront (Leven, 1968), the majority of the relevant studies acknowledge the possibility of multiple optima in city size (Arnott, 1979). Multiple optima can be attained, given that each city is endowed with a different level of external economies to exploit, depending on its economic base. Such varied scopes of external economies across cities arise because each city is specialized in a small but distinct set of industrial sectors and the extent of external economies of scale substantially differs across sectors (Henderson, 1974).

The possibility of multiple optima suggests a serious challenge in application of this approach to build a forecast model that covers a long time frame and a broad geographical space. The primary reason for the challenge is that given the existence of multiple optima, researchers need to describe each city's economic structure accurately. This is a demanding task if the goal is to cover large areas that include many cities, as essentially, each city must be modeled independently. Related data requirements are another challenge to be cleared for that purpose. In addition, predictive analysis of the optimization results requires many economic variables that may be impossible to forecast with any reliability. Thus, the approach does not offer much real predictive power.

There may be other challenges that limit application of this approach. Varied perspectives surrounding the optimality concept may be one of them. While early literature defined optimality in terms of net-costs of public services associated with city size (Richardson, 1972), more recent studies tended to define optimal city size as a population level where external economies of scale are maximized or the marginal social utility of a unit population increase is close to zero (Henderson, 1988). In terms of objective functions, many have adopted certain forms of net marginal benefit function defined on a particular city size and its unit increase, despite some differences in detail. However, there has been controversy regarding from whose perspectives such functions are to be optimized. Some (Yezer and Goldfarb, 1978; Capello and Camagni, 2000; Zheng, 2007) focus on optimality from a city's own perspective; others (Suh, 1991; Henderson, 2000; Venables, 2005; Au and Henderson, 2006) see the need to embrace a regional or national perspective.

3. DATA AND METHOD

As reviewed in the previous section, the existing approaches impose serious limitations on application to downscaling global population. An approach built on Zipf's law can lead to a descriptive model having merit in covering large samples, but the lack of theoretical foundations limits the model's capability of predictive analysis. On the other hand, an approach built on the optimization framework has limited applicability mainly due to intensive data requirements, although it can provide a useful tool for both descriptive and predictive analyses. Our approach combines aspects of the two approaches, estimating distribution or regularities in population density and conditioning those on economic variables that are widely available.

3.1 Data

This study uses 0.25°×0.25° global population density data sets, developed by the Socioeconomic Data and Applications Center (SEDAC) at Columbia University. Our decision to use grid cell data, instead of conventional population data, is to avoid the arbitrariness of administrative definitions. We continue to use the word "city" or "urban area" to describe our basic unit of analysis, but what we mean is population located in a geographically defined grid of 0.25°×0.25° with at least 2 persons/km². Using grid cell data is also consistent with much environmental research that has motivated this analysis, and prevents the robustness of our results from being dependent on an artificially chosen sample size, which has been a serious critique leveled against many empirical works on rank-size regularities.

At present, SEDAC distributes the data set for six time points from 1990 to 2015 at a five-year interval, but we use the data set for the first four time points between 1990 and 2005, benchmarked with actual national census data. In addition, our analysis focuses only on 112,080 grid cells or 44% of the world's total land area coverage, leaving out rural or uninhabitable land area that have population density less than 2 persons/km². Note from **Figure 1** that this definition encompasses much area that would be considered rural, but here we are focusing on broader geographic distribution of population.

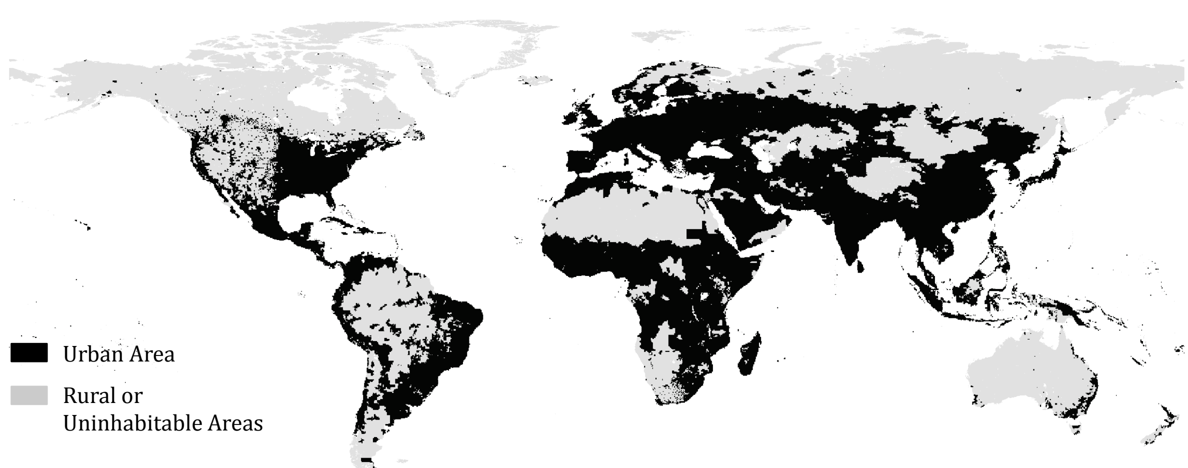

Figure 1. Grid Cells Defined as Inhabitable Land. Source: Created from SEDAC (2005).

In addition to the population data, our analysis requires additional inputs that explain country-specific natural endowments, development stage, and industrial structure. Most data related to these variables are from the World Development Indicator Database, published by the World Bank (World Bank, 2011).

3.2 Conceptual Lens

As stated earlier, our hypothesis is that spatial distribution of global population can be explained by the combination of natural regularity inherent in city size and rank, as well as the socio-economic characteristics of the country or region in which the city is located. Our method and empirical analysis test this hypothesis.

Our method first identifies which country each of the 0.25°×0.25° global population density grid cells belongs to, and sorts all of the grid cells belonging to each country by their population density. Then, we estimate a Beta distribution function, which allows substantial flexibility in distributional shape, for each country, using the global population data for four time points. Third, cross-sectional and time-series analyses are conducted to estimate the shape parameters of a Beta distribution as functions of macroeconomic variables for each country, such as per capita income or manufacturing share of the national income. The final step is to link estimated Beta distributions with the rank-size rule.

3.3 Rank-Size Relationship and Beta Distribution Function

A fundamental step in our method is to link each grid cell's density level with its rank in a national urban hierarchy. It is straightforward to sort grid cells of one country by population size (or density level) and endow each grid cell with a particular rank, but it is not as straightforward to estimate a particular cell's population size (or density level) from its given rank. That is, if we know only each grid cell's rank without having information on its density level, estimating the latter by using the former would need an additional assumption and a special mathematical setup, introduced below. The assumption necessary to make size estimation from rank feasible is that the rank endowed to a particular cell within an urban hierarchy, at either national or regional levels, is stable across time periods. This stability assumption allows us to apply the information on city-size distribution of a region, acquired from existing data, to the past or future time frames, for which reasonably downscaled population data sets are not available. When the stability assumption is adopted, then we can translate the rank into the size (density level) by the following mathematical equations and transformation.

Note that the rank of a grid cell having a density level of x (R_x) and the cumulative probability corresponding to that density level (cdf(x)) satisfies the equation:

$$R_x = T\left(1 - \mathrm{cdf}(x)\right) + 1 \tag{1}$$

where T stands for a total number of grid cells in a country. Equation 1 can be transformed into the following equation:

$$\mathrm{cdf}(x) = 1 - \frac{R_x - 1}{T} = y \tag{2}$$

Then, the inverse transformation of Equation 2 will give us the following:

$$x = \mathrm{cdf}^{-1}(y) \tag{3}$$

Equation 3 tells us that we can estimate the density level of a particular grid cell (x), as long as we know the cell's rank (R_x), the total number of grid cells in a country (T), and the relevant cumulative distribution function (CDF). As stability in rank is assumed and the sample size for our analysis (i.e., grid cells defined as inhabitable land) is determined from the 1990-2005 data set, we already know the first two parameters R_x and T. Then, arriving at reasonable estimates for cdf(x) would complete this given rank-size transformation process.

Adopting the approach in Asadoorian (2008), we fit the distribution of grid cell-based population density data to the Beta distribution function. The Beta probability distribution function (PDF) is given in the following form:

$$f(X) = \frac{X^{\alpha-1}(1-X)^{\beta-1}}{\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt} \tag{4}$$

where X is a random variable having values between 0 and 1, and both α and β are parameters that jointly determine the shape of the PDF. As a Beta PDF is defined on a random variable taking values between 0 and 1, we normalized each cell's density value to the maximum grid-cell value of a country after taking the natural logarithm. The Beta PDF can easily be transformed into the following CDF form:

$$F(X) = \frac{\int_0^X t^{\alpha-1}(1-t)^{\beta-1}\,dt}{\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt} \tag{5}$$

The primary reason we fit the given rank-size distribution to the Beta distribution function is to take advantage of the function's flexibility of shape. It is likely that our grid cell-based data exhibit tendencies different from those in existing theories or empirical findings, as most literature on city size distribution is based on cities as defined in political or economic terms. As shown in **Figure 2**, density grid cells do not necessarily follow log-normal distributions, as expected from the application of Gibrat's law or a proportionate urban growth process. Instead, the histograms shown in the figure demonstrate that the distribution of the log of population density can be skewed substantially from the mean density level either left- or rightward.

The distributional patterns differ not only across regions but also across time periods, suggesting that the distribution of urban population density is dynamic and actively interacts with socio-economic conditions. The greatest advantage of a Beta PDF for this research project is that it can illustrate various distributional patterns, depending on its two shape parameters (Nelson and Preckel, 1989). This flexibility allows us to substantially relax the constraint associated with a certain, presupposed distributional form, such as the proportionate growth assumption attached to the log-normal distribution of city size. In addition to the robustness of rank-size relationship against the definition of a city, which we secure by using grid-cell data, this relaxed constraint in presupposed city-size distributional shape is another main improvement that our method can bring to the conventional approaches.

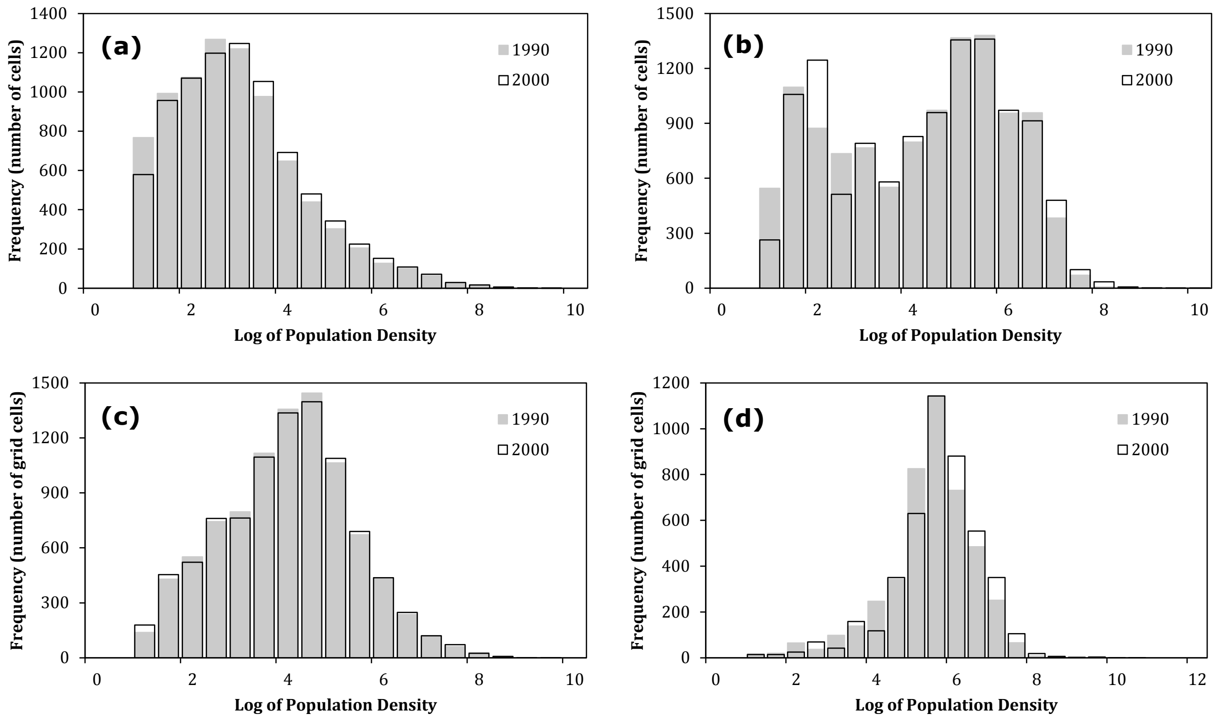

Figure 2. Histogram of Density Grid Cells, 1990 and 2000: **(a)** USA, **(b)** China, **(c)** Europe Union, **(d)** India. Source: Created from SEDAC (2005).

3.4 Beta Distribution as Function of Socio-economic Variables

Once Beta CDFs are estimated by country and time, then we analyze the empirical relationship between the estimated shape parameters and several socio-economic variables. The hypothesis underlying this regression analysis is that the city size distribution of a society depends highly on a certain set of socio-economic conditions that the society confronts, which distinguishes our method from the conventional approaches built on empirical rank-size regularities. If this hypothesis holds, then our model, though also relying on the rank-size relationship, would have substantially improved forecasting power of spatial distribution of population, compared with traditional models. In detail, we use the following two equations to build empirical relationships between the two shape parameters of Beta CDF and a set of independent variables. In the equations, where $\boldsymbol{\gamma}$, $\boldsymbol{\pi}$, $\boldsymbol{\delta}$, and $\boldsymbol{\eta}$ are coefficient vectors and ε_i and υ_i are error terms, the two shape parameters of a Beta CDF for region i (α_i and β_i) are estimated by using $\mathbf{X}$ and $\mathbf{D}$, which denote vectors of experimental and dummy variables, respectively.

$$\ln(\alpha_i) = \boldsymbol{\gamma}'\ln(\mathbf{X})_i + \boldsymbol{\pi}'\mathbf{D}_i + (\text{interaction terms}) + \varepsilon_i \quad (6)$$

$$\ln(\beta_i) = \boldsymbol{\delta}'\ln(\mathbf{X})_i + \boldsymbol{\eta}'\mathbf{D}_i + (\text{interaction terms}) + \upsilon_i \quad (7)$$

We theorize that the geographic distribution of population—or the variation in parameters of the Beta distributions—is explained by national or regional characteristics, including natural endowments, industrial structure, development stage, political centrality, and historical or cultural differences. Physical geographic features or natural endowments tend to determine basic patterns of population distribution, and the relative importance of agriculture or industry can significantly modify such basic patterns of demographic geography (Davis and Weinstein, 2002). It is also widely believed that a certain development stage that involves rapid growth of the market or urban sectors leads to massive inter-city migration of domestic population and affects the geographic distribution of population (Moomaw and Shatter, 1996). Political centrality or dictatorship is known to be positively correlated to the primacy of the largest city in national urban hierarchies and to affect the Pareto exponent of Zipf's plot (Rosen and Resnick, 1980; Ades and Glaeser, 1995). We assume that a portion of demographic geography, left unexplained by these four categories, reflects inter-regional or inter-temporal variations in culture and history. We quantify these theoretical considerations as shown in **Table 1**. In sum, our regression models include a total of nine explanatory variables that describe national or regional characteristics, and multiple sets of dummy variables and interaction terms that consider region- or time-specific effects on intercepts and slopes.

Table 1. Variables Used in Regression Analysis.

Category	Variables	Notations
+ Natural Endowments	Population density in total land area (persons/km²)	X_1
	Population density in urban area only (persons/km²)	X_2
	% share of arable land in total land area	X_3
	Arable land area (km²)	X_4
+ Industrial Structure	% share of agriculture in GDP	X_5
	% share of manufacturing in GDP	X_6
+ Development Stage	GDP per capital in PPP terms (constant international $)	X_7
+ Political Centrality	Size of the highest density cell in each country or region	X_8
	Primacy of the highest density area in each country	X_9
+ Region-specific Fixed Effects (Region Dummies)	1 if belonging to the Americas or Oceania; 0 otherwise.	D_1
	1 if belonging to Europe (including Russia); 0 otherwise.	D_2
	1 if belonging to Asia; 0 otherwise.	D_3
	1 if belonging to Africa; 0 otherwise.	D_4
+ Time-specific Fixed Effects (Year Dummies)	1 if year 1995; 0 otherwise.	D_5
	1 if year 2000; 0 otherwise.	D_6
	1 if year 2005; 0 otherwise.	D_7
+ Interaction Terms	Combinations of log of X_i and regional dummies	$D_i \cdot \ln X_i$

4. RESULTS

4.1 Goodness of Fit

Our empirical analysis focuses on four time points (1990, 1995, 2000, and 2005) and 65 countries, a brief description of which is shown in **Table 2**. Those countries included in our sample are chosen based on the availability of reliable time series statistics on the independent variables of our primary interest and on the size of grid cell coverage (cutoff of 300 grid cells) to ensure good fits of actual distribution to the Beta function. The actual distribution of population density in all countries included in our sample fits well to the Beta CDF. For example, the goodness of fit for each of the 65 countries in 1990, when measured in R^2 terms, is shown to be 94% or higher (**Table 3**). Estimated Beta CDFs, compared with actual CDFs, for selected cases in 1990 are displayed in **Figure 3**. These goodness-of-fit measures suggest that the Beta distribution function provides a good platform that can be used as an appropriate functional form for the distribution of global population density.

Table 2. Summary of Sample Size.

	Number of Countries	Total Grid Cell Coverage (cells)	Average Grid Cell Coverage by Country (cells)
By Continent			
+ Asia	13	27,725	2,133
+ Europe*	18	23,972	1,332
+ Africa	22	20,027	910
+ New Continent**	12	25,329	2,112
By Income Group* **			
+ Advanced Economies	12	17,302	1,442
+ Emerging/Developing Economies	53	79,751	1,505

* Includes Russia and other member countries of the former Soviet Union; ** Includes North and South Americas and Oceania; *** Adopts the country group classification by the International Monetary Fund.

Table 3. Summary of Goodness of Fit, 1990.

	Total	$0.98<R^2\leq1$	$0.96<R^2\leq0.98$	$0.94<R^2\leq0.96$
Number of Observations	65	51	12	2

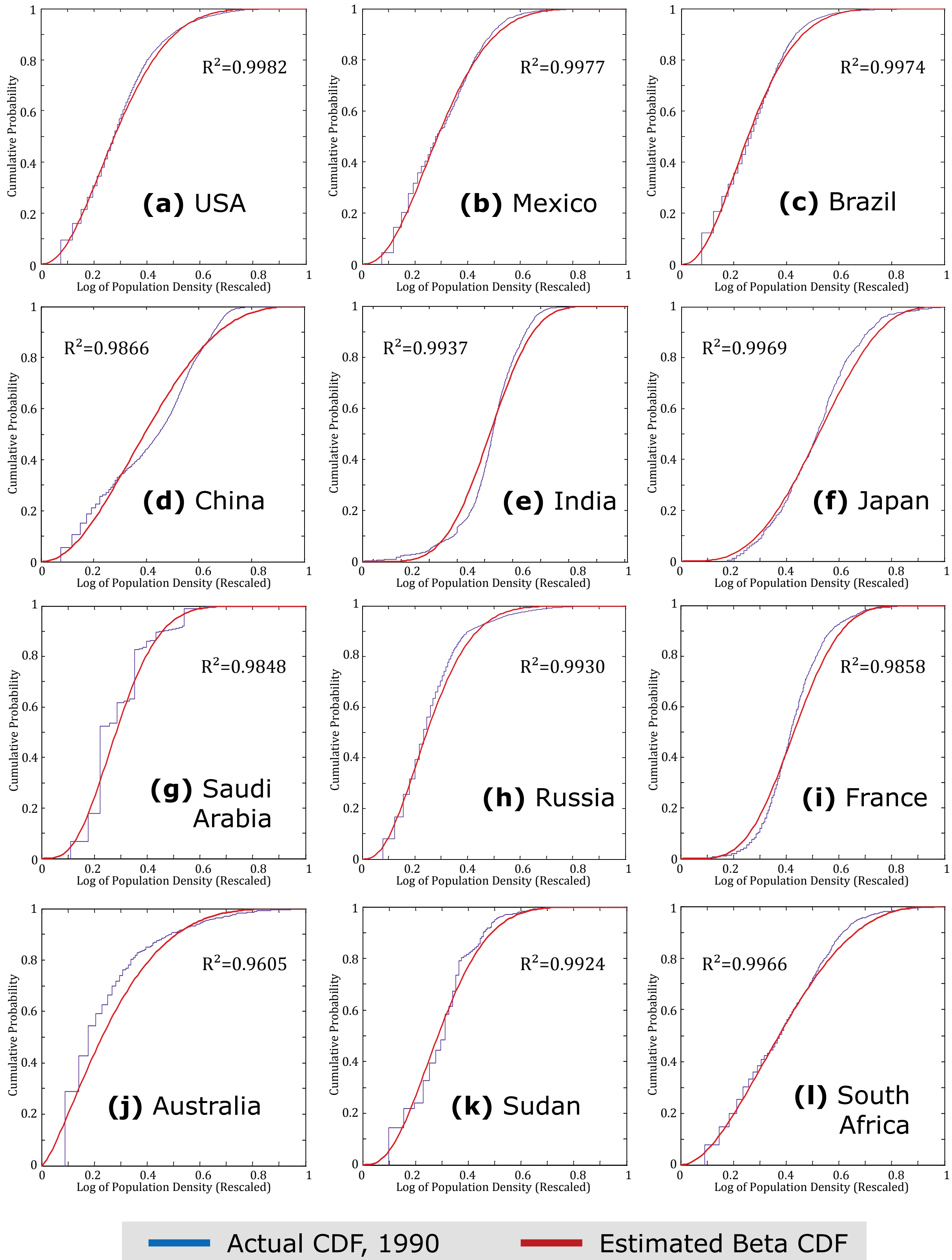

Figure 3. Actual and Estimated CDFs for Selected Countries, 1990.

4.2 Socio-economic Variables as Determinants of Beta CDF

Our next question is whether the distributional pattern of each country's population can be explained to a substantial degree by key geographic and socioeconomic variables or not. This analysis attempts to link empirical rank-size relationships and the urban economics literature so that our model can strengthen its explanatory and forecasting power for urban growth patterns.

Our regression analysis shows that the two shape parameters of the given Beta distribution function are well accounted for by a subset of socio-economic variables listed in Table 1 (**Table 4**). Population density (X_1 and X_2), arable land coverage (X_3), gross domestic product (GDP) per capita (X_7) are common explanatory variables for both shape parameters, while the manufacturing share of GDP (X_6) and the size of maximum density grid cell (X_8) are good predictors of only one of the two parameters. In addition, our analysis finds varied marginal effects, involving changes in slopes, of per capita GDP on the shape parameters on the New Continent (D_1), Asia (D_3), or Africa (D_4), and a region-specific fixed effect, involving changes in intercepts, for Asian countries. The regression models for both shape parameters are significant at the 1% level, and have high explanatory power. The models for the shape parameters α and β have R^2 values of 62% and 71%, respectively.

Table 4. Regression Analysis Results.

LHV†	ln(α)		ln(β)	
RHV‡	**Coefficient**	**Standard Error**	**Coefficient**	**Standard Error**
lnX_1	0.215**	0.041	0.117**	0.032
lnX_2	-0.260**	0.053	-0.637**	0.041
X_3§	0.025**	0.002	0.019**	0.002
lnX_6	-0.150**	0.044	(excluded)	
lnX_7	-0.085**	0.024	-0.041**	0.014
lnX_8	(excluded)		0.272**	0.016
D_3	-1.385**	0.408	-1.080**	0.280
D_1·lnX_7	(excluded)		0.011*	0.005
D_3·lnX_7	0.177**	0.047	0.148**	0.034
D_4·lnX_7	-0.025**	0.008	(excluded)	
(Constant)	2.260**	0.210	1.563**	0.119
R^2	0.620		0.709	
Observations	260		260	

† Left hand side variable; ‡ Right hand side variable; § No logarithm is taken for Z_3; * Statistically significant at 5% level; ** Statistically significant at 1% level.

4.3 Comparison of Actual and Estimated Population Growth

To test the robustness of our model, we apply our model to the 2000 data set, and compare our estimates with actual grid-specific density levels. The first stage of our robustness test involves a comparison of two sets of the shape parameters of a Beta CDF for each of the 65 countries—one set estimated from actual density data and the other set estimated from the regression models shown in Table 3. Overall, Beta CDFs estimated from our regression models are reasonably close to those estimated from actual density data, presenting 78% and 65% of R^2 values for α and β, respectively (**Figure 4**).

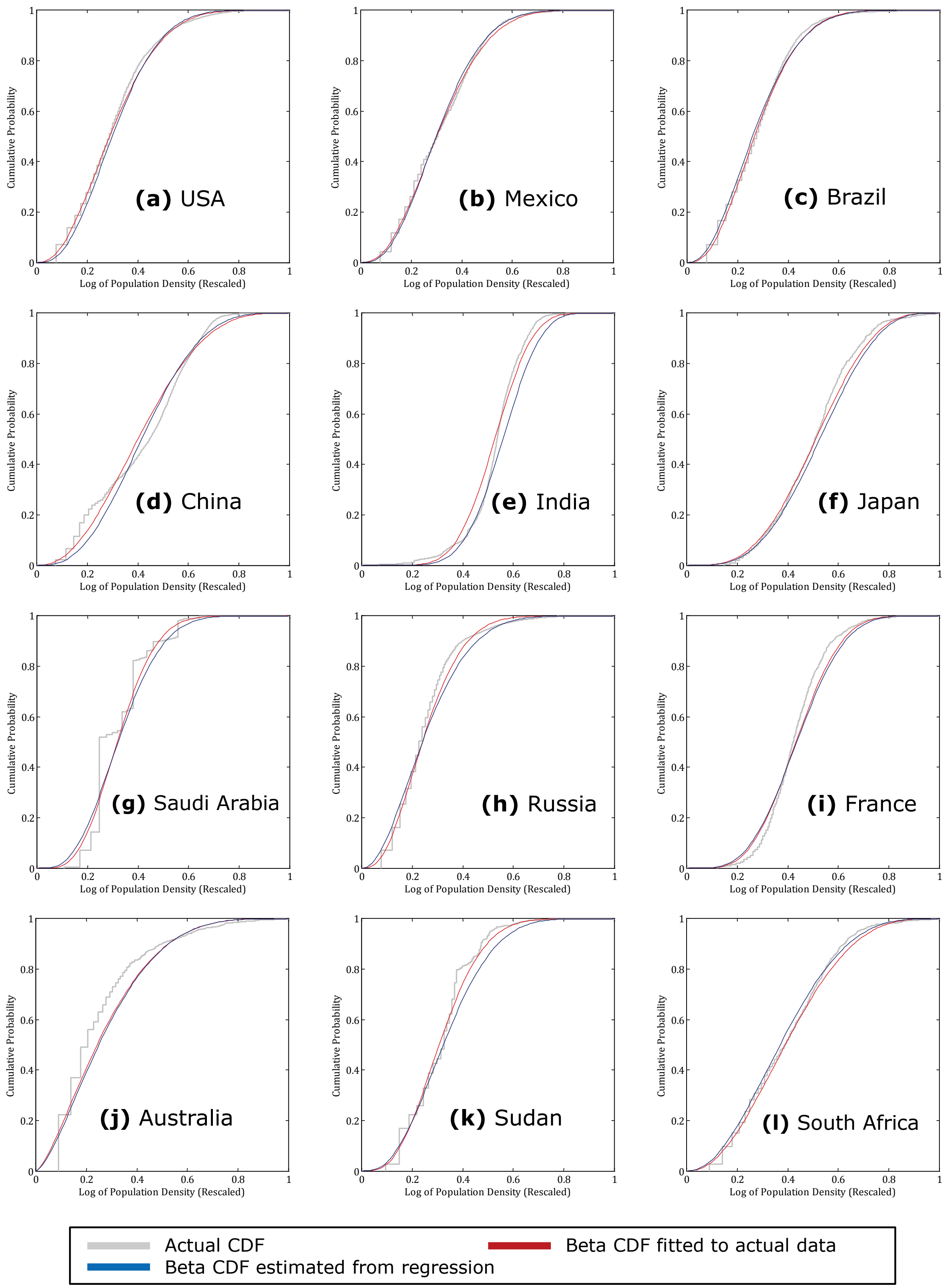

Figure 4. Beta CDFs Estimated from Regression, Compared with Actual CDFs and Beta CDFs fitted to Actual Data, for Selected Countries, 2000.

The second stage of our test focuses on the accuracy with which our model can generate grid cell-specific density estimates. This stage of the test aims to cover the entire sample of grid cells defined as urban area, but such coverage requires an additional task—grouping countries into sizable regions. This regional aggregation is to ensure a large number of cells for each region, so that sorting grid cells by density and estimating related CDFs by region are feasible. We group the entire world into 16 regions, with reference to the aggregation scheme adopted for the fifth version of the Emissions Prediction and Policy Analysis (EPPA5) model[1] (**Figure 5**). Although other regional aggregations are possible, our regional aggregation scheme benchmarks EPPA5 to use the model's macroeconomic simulation results as input for long-term projection of grid cell-specific density levels (see Section 5).

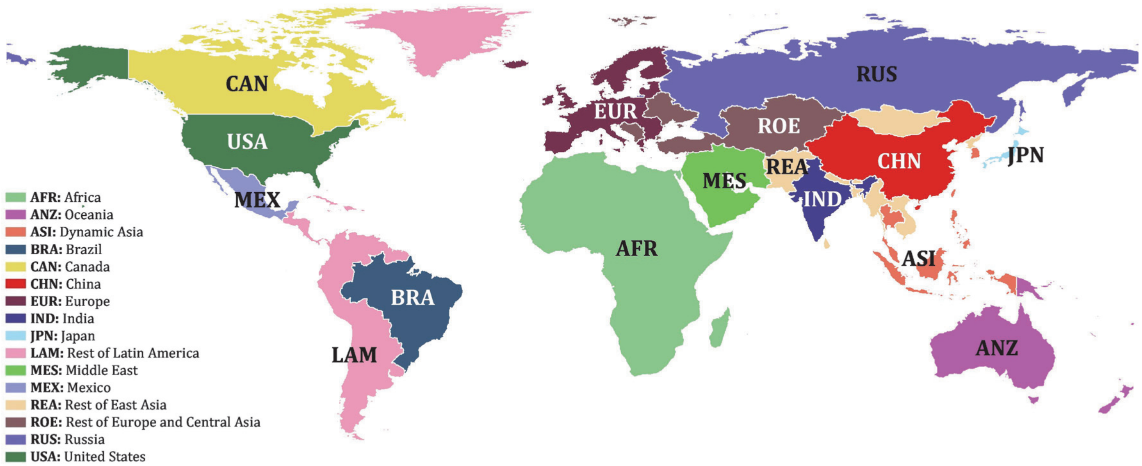

Figure 5. Regional Aggregations in EPPA5.

Our second-stage test results show that our model is capable of forecasting cell-specific density levels to a substantially accurate degree. When actual 2000 density data and our estimates are compared in a cell-by-cell fashion, our model explains 76% of the total sum of squares. Accuracy in the case of short-run projection may further increase through the use of information on cell-specific errors computed from previous years' data sets. We adjust the 2000 density estimates with cell-specific errors from the 1990 data set, as shown in the following equation, where $y_{i,0}$.denote actual population density for grid cell *i* in the base year (1990, in our exercise), and $\hat{y}_{i,t}$ and $\hat{y}^*_{i,t}$ refer to unadjusted and adjusted density estimates for grid cell *i* in time *t*, respectively.

$$\hat{y}^*_{i,t} = \hat{y}_{i,t} + (y_{i,0} - \hat{y}_{i,0}) \tag{8}$$

R^2 of adjusted estimates for 2000 reached over 0.98 (R^2 of 0.76 for unadjusted estimates). **Figure 6** displays actual cell-specific population growth rates between 1990 and 2000, compared with our unadjusted and adjusted estimates.

[1] EPPA5 is a multi-region, recursive dynamic computable general equilibrium (CGE) model built on the GTAP7 dataset. Paltsev *et al*. (2005) offers a detailed description of its earlier version.

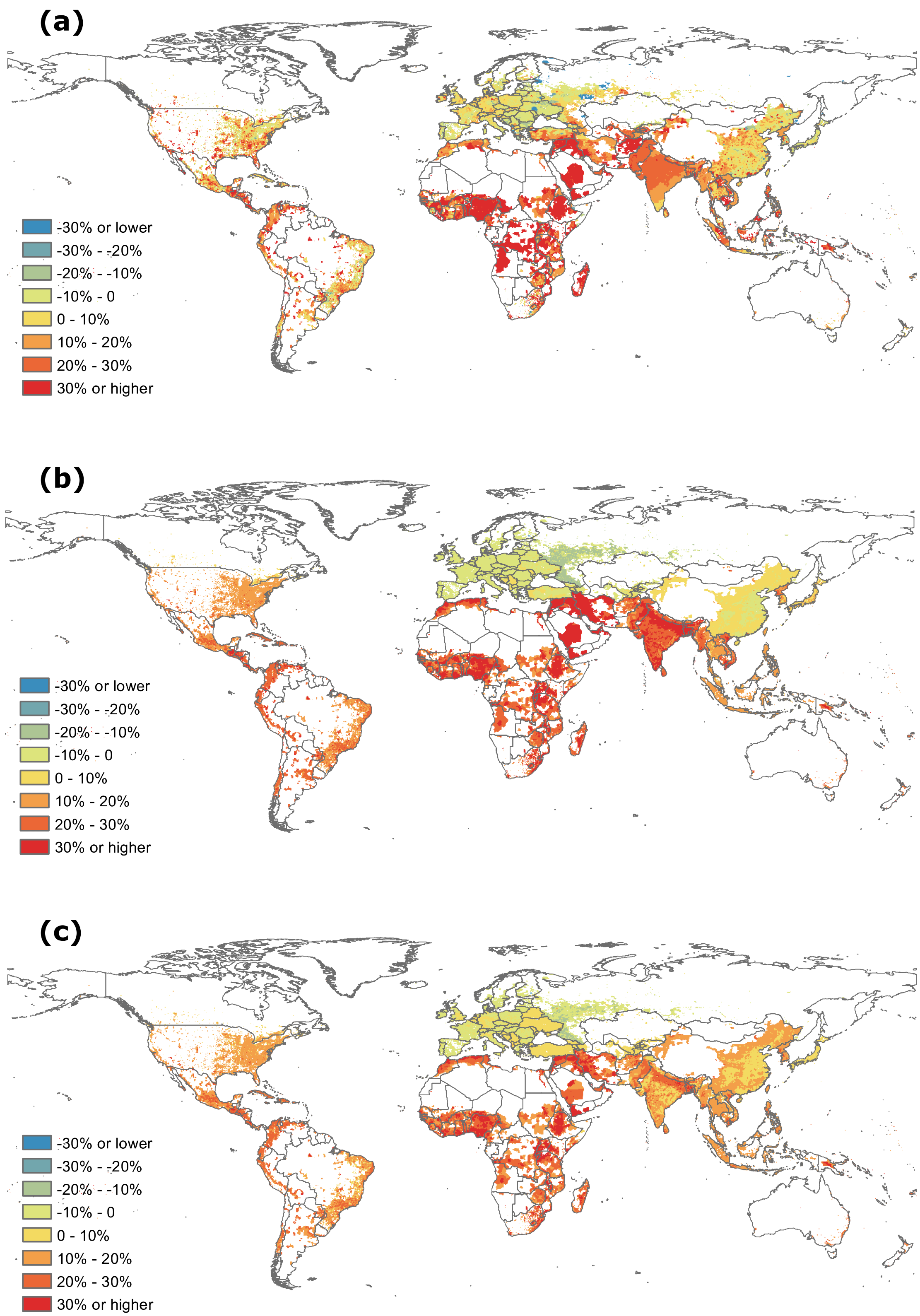

Figure 6. Population Growth Rates, 1990-2000: **(a)** Actual Data, **(b)** Unadjusted Estimates, **(c)** Adjusted Estimates. Measured in % change of the 1990 population level. Only the grid cells, whose density levels in 2000 were 10 persons/km² or higher, are displayed.

5. LONG-TERM PROJECTIONS

In this section, we aim to show how our method can be extended for predictive analysis of the spatial distribution of global population. For this purpose, we apply our model to up to the year 2100, in connection with EPPA5, and present how the patterns of spatial distribution of global population look in the future.

5.1 Prime Grid Cells and Other Inputs

The application of our population distribution model for future time periods requires that the independent variables listed in Table 3 be estimated for the same time periods. As done in Section 4.3, we apply our model to the 16 global regions used for EPPA5, so that we can use the data simulated by EPPA5 as inputs for our model. We assume that the geographic conditions such as total land area, urban land area, and arable land area are unchanged. We depend for future population data on United Nations (2010), which provides country-specific population projections up to the year 2300. In the case of the two proxy variables for economic structure (X_6) and development stage (X_7), we depend on the simulated results from EPPA5. Density levels for each region's prime grid cell (X_8) are estimated by applying city-specific growth rates computed from the United Nations data set for the world's large urban agglomerations from 1950 to 2025 (United Nations, 2009). The 2021-2025 growth rates are applied for the time period later than 2025.

5.2 Projected Global Population Distribution

The results of our predictive analysis show that each region is expected to face a spatial distribution of population density in 2100, which differs substantially from the one in 2000 (**Figure 7**). In some regions, further urbanization or population centralization is expected. Most notably, Africa would see the emergence of a number of densely populated urban areas, whose density levels go beyond 1000 persons/km², and the thinning of sparsely populated areas, whose density levels are 10 persons/km² or lower. Although the change is less dramatic than in Africa, other regions like Europe, the United States, and Oceania also show similar changing patterns of demographic geography. On the other hand, other regions, such as Brazil and Russia, would encounter with an increasingly decentralized population distribution pattern in the future.

China and India—the two most populous countries—are forecasted to experience urban growth patterns dissimilar to one another. In the future, China's population is expected to distribute in a more decentralized way, leading to fewer very densely populated grid cells. By contrast, in India the existing demographic geography is likely to be somewhat strengthened with a gradual urbanization process still in progress.

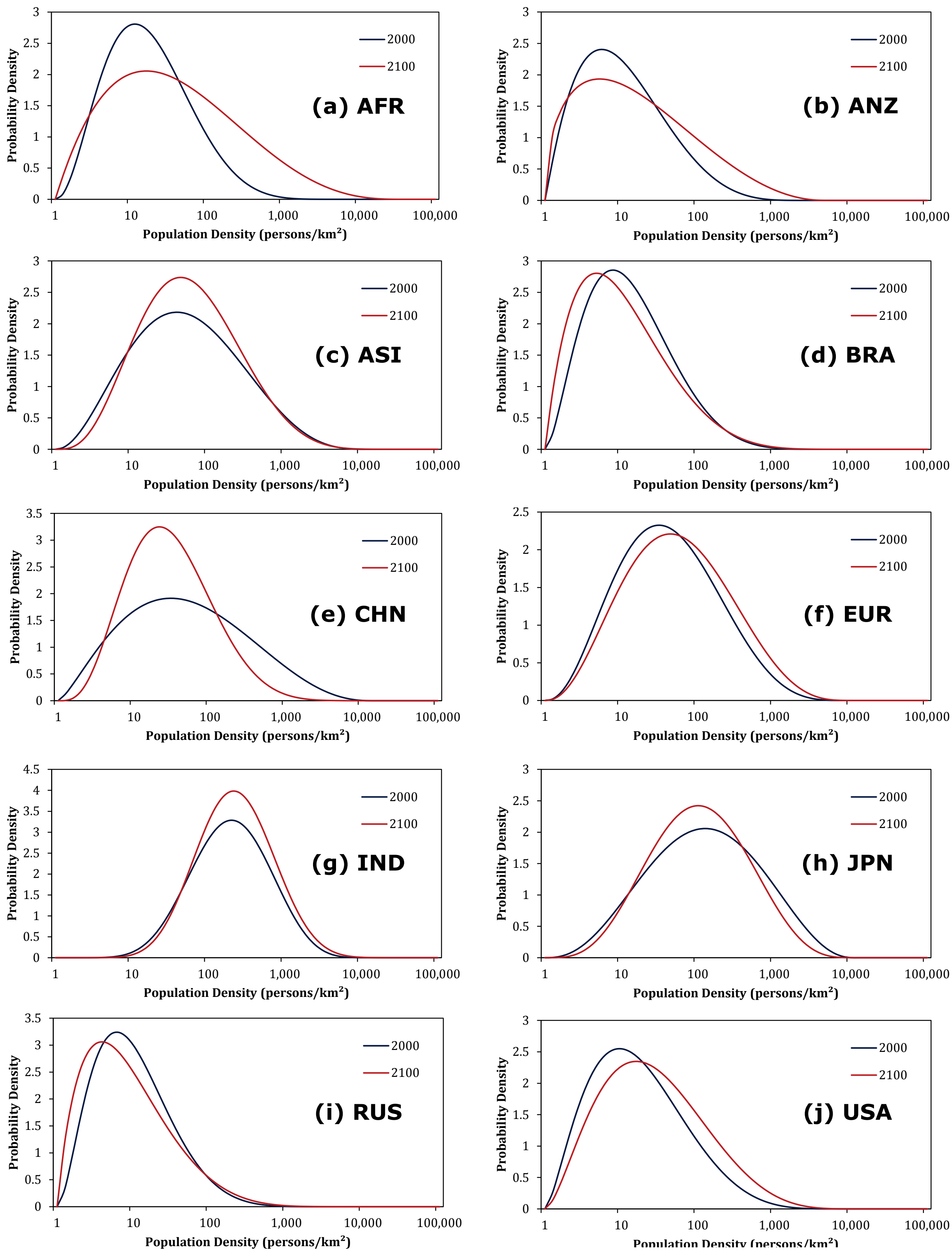

Figure 7. Probability Distribution of Population Density for Selected EPPA5 Regions, 2000 and 2100. *Y*-axis refers to the probability density, where an estimated Beta CDF is evaluated at the log of *X* rescaled with its regional maximum 1.

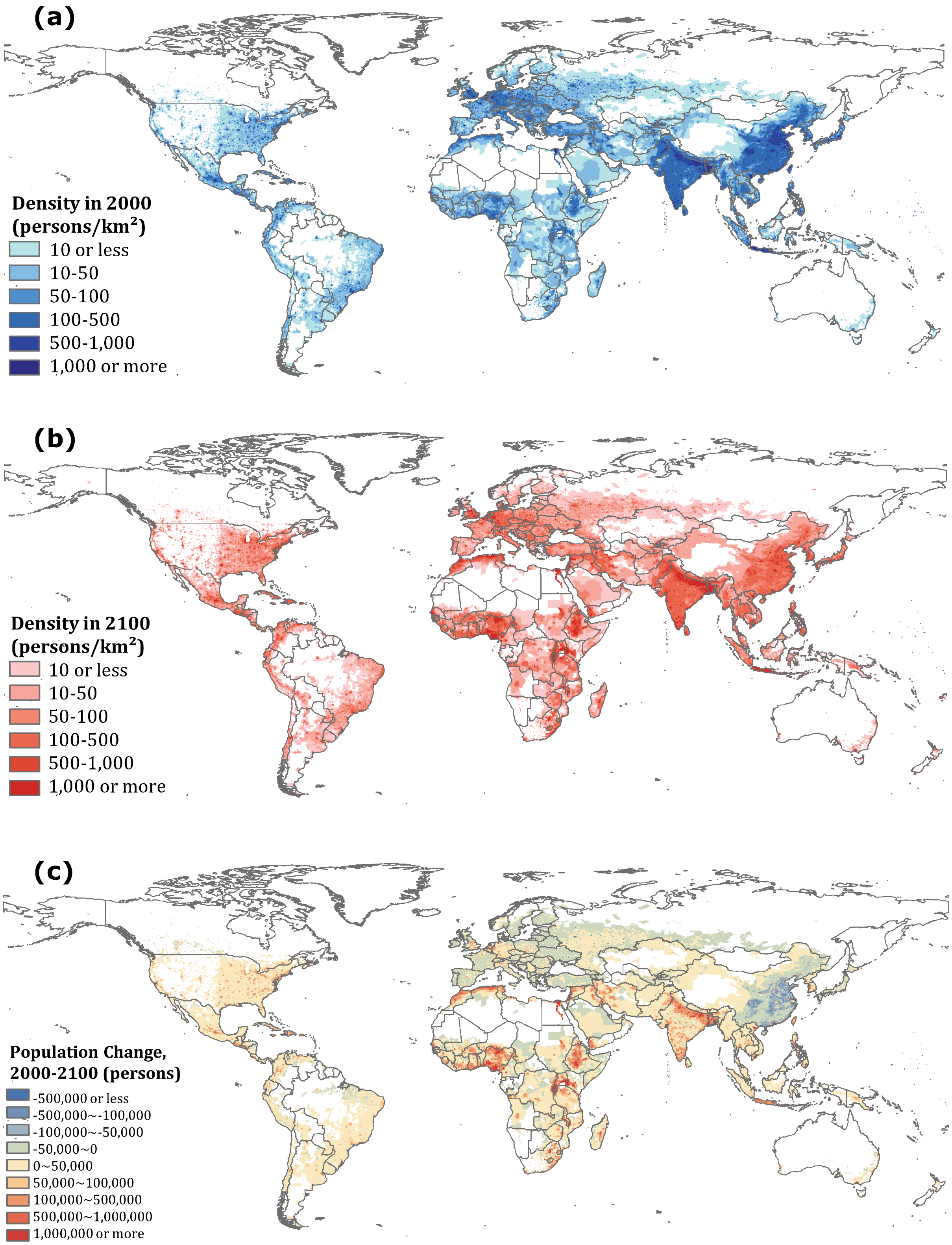

Figure 8. Comparison of Global Population Distribution between 2000 and 2100: **(a)** Actual Population Density in 2000; **(b)** Estimated Population Density in 2100 (unadjusted estimates); **(c)** Population Change between 2000 and 2100.

Figure 8 compares cell-specific global population distribution between 2000 and 2100. Africa, India, and the Middle East are at the center of the future global population growth, and population growth in these regions is led by several urban agglomerations. In Africa, such urban agglomerations include the Casablanca-Algiers and the Cairo-Alexandria corridors of the North, the Lagos-Ibadan corridor and the Port Harcourt-Benin City-Enugu triangle of the West, and the Addis Ababa metropolitan area and the Kigali-Kampala-Nairobi corridor of the East. In India, the Delhi metropolitan area, the Varanas-Patna corridor, and the Kolkata-Dhaka corridor are expected to lead the nation's population growth. In the Middle East, the growth of the Amman-Beirut-Hamah corridor and the Baghdad and Tehran metropolitan areas is distinguished.

In the other global regions, a relatively modest range of population change is anticipated with few exceptions, such as the Jakarta-Bandung-Semarang corridor of Indonesia's Java Island. Most of China's main urban areas are forecasted to experience substantial decreases in population.

6. CONCLUSIONS

In this paper, we introduce a new approach that can be used to downscale aggregated global population into fine grid cell levels. The originality of our model lies in its attempt to link empirical rank-size regularities with the socio-economic driving forces underlying them. The linkage of these two aspects is intended to reflect the cross-regional and time variations in city-size distributions, which we have identified and demonstrated through empirical analysis in Sections 3 and 4. Also, we propose to use grid cell population density data, instead of the data constructed on politically or socioeconomically defined urban boundaries, and thereby avoid defining a city in a discretionary fashion. Accordingly, our method developed in this study is not subject to a few problems apparent in conventional approaches, such as city-size distributions affected by the sample size (or a cutoff population level to be considered a city) and the predetermined list of cities excluding the possibility of the emergence of new urban agglomerations in the future.

Our main contribution to the urban studies literature is twofold. First, our study challenges the conventional view that rank-size regularities support the random growth hypothesis, and thus contributes to expanding related academic debates within the urban studies circle. As reviewed in Section 2, Zipf's rank-size rule has often been translated as empirical evidence of the proportionate urban growth process, where a city's growth rate is independent of its size. So far, multiple studies have shown that rank-size regularities and the random growth hypothesis coincide with each other and support a log-normal distribution of city size. Our study, however, shows that when avoiding an artificial definition of a city, the city-size distribution of a region can differ significantly from a log-normal distributional form and can be skewed substantially either left- or rightward. We also demonstrate that the city-size distributional form of a region not only differs from that of other regions, but also can change substantially over time, depending on socioeconomic conditions that the region faces. All of these findings undermine the random growth hypothesis.

The other contribution to the field can be that our study introduces a downscaling model, which can be applied to all the global regions, including the regions for which reliable subregional or subnational time-series population data sets are limitedly available. We have also shown that our model can be extended easily to function as a forecasting model, in connection with other economic models such as EPPA5. As discussed in Section 2, existing approaches based on the rank-size rule associated with the random growth hypothesis or the economic optimization framework lack applicability to construct forecast or downscaling models comparable to ours, mainly due to their intensive data requirements or their failure to associate the empirical regularities with socioeconomic conditions underlying such regularities. We believe that our study can pave the way for various academic or practical research projects, which need spatial distribution of global population at fine grid cell levels as key input.

Acknowledgments
The Joint Program on the Science and Policy of Global Change is funded by the U.S. Department of Energy, Office of Science under grants DE-FG02-94ER61937, DE-FG02-93ER61677, DE-FG02-08ER64597, and DE-SC0003906; the U.S. Department of Energy, National Renewable Energy Laboratory under grant XEU-0-9920-01; the U.S. Environmental Protection Agency under grants XA-83344601-0, XA-83240101, PI-83412601-0, and RD-83427901-0; the U.S. National Science Foundation under grants SES-0825915, EFRI-0835414, BCS-0410344, ATM-0329759, DMS-0426845, and AGS-0944121; the U.S. National Aeronautics and Space Administration under grants NNX07AI49G, NNX08AY59A, NNX06AC30A, NNX09AK26G, NNX08AL73G, NNX09AI26G, NNG04GJ80G, NNG04GP30G, and NNA06CN09A; the U.S. National Oceanic and Atmospheric Administration under grant NA070AR4310050; the U.S. Federal Aviation Administration under grants 06-C-NE-MIT and 09-C-NE-MIT; the U.S. Department of Transportation under grant DTRT57-10-C-10015; the U.S. Department of Agriculture under grant 58-0111-9-001; the Electric Power Research Institute under grant EP-P32616/C15124; and a consortium of 40 industrial and foundation sponsors (for the complete list see http://globalchange.mit.edu/sponsors/current.html).

7. REFERENCES

Ades, A.F., and E.L. Glaeser, 1995: Trade and Circuses: Explaining Urban Giants. *Quarterly Journal of Economics*, **110**(1):195-227.
Arnott, R., 1979: Optimal City Size in a Spatial Economy. *Journal of Urban Economics*, **6**:65-89.
Asadoorian, M.O., 2008: Simulating the Spatial Distribution of Population and Emissions to 2100. *Environmental and Resource Economics*, **39**(3):199-221.
Au, C.-C., and J.V. Henderson, 2006: Are Chinese Cities Too Small? *Review of Economic Studies*, **73**(3):549-576.
Black, D., and J.V. Henderson, 2003: Urban Evolution in the USA. *Journal of Economic Geography*, **3**(4):343-372.
Capello, R., and R. Camagni, 2000: Beyond Optimal City Size: An Evaluation of Alternative Urban Growth Patterns. *Urban Studies*, **37**(9):1479-1496.
Carroll, G.R., 1982: National City-Size Distribution: What Do We Know After 67 Years of

Research? *Progress in Human Geography*, **6**:1-43.
Córdoba, J.-C., 2008: On the Distribution of City Sizes. *Journal of Urban Economics*, **63**:177-197.
Davis, D.R., and D.E. Weinstein, 2002: Bones, Bombs, and Break Points: The Geography of Economic Activity. *American Economic Review*, **92**(5):1269-1289.
Eeckhout, J., 2004: Gibrat's Law for (All) Cities. *American Economic Review*, **94**(5):1429-1451.
Fujita, M., P. Krugman, and A.J. Venables., 1999: *The Spatial Economy: Cities, Regions, and International Trade*. MIT Press: Cambridge, MA.
Gabaix, X., 1999: Zipf's Law for Cities: An Explanation. *Quarterly Journal of Economics*, **114**(3):739-767.
Gabaix, X., and Y.M. Ioannides, 2004: The Evolution of City Size Distributions. In *Handbook of Regional and Urban Economics, Volume 4: Cities and Geography*, edited by J. V. Henderson and J.-F. Thisse. Elsevier Science: New York, NY.
Giesen, K., A. Zimmermann, and J. Suedekum, 2010: The Size Distribution across All Cities: Double Pareto Lognormal Strikes. *Journal of Urban Economics*, **68**:129-137.
González-Val, R., 2010: The Evolution of U.S. City Size Distribution from a Long-term Perspective (1900-2000). *Journal of Regional Science*, **20**:1-21.
Guérin-Pace, F., 1995: Rank-Size Distribution and the Process of Urban Growth. *Urban Studies*, **32**(3):551-562.
Henderson, J.V., 1974: The Sizes and Types of Cities. *American Economic Review*, **64**:640-656.
Henderson, J.V., 1988: *Urban Development: Theory, Fact, and Illusion*. Oxford University Press: New York, NY.
Ioannides, Y.M., and H.G. Overman, 2003: Zipf's Law for Cities: An Empirical Examination. *Regional Science and Urban Economics*, **33**:127-137.
Leven, C.L., 1968: Determinants of the Size and Spatial Form of Urban Areas. *Papers of the Regional Science Association*, **22**:7-28.
Malacarne, L.C., R.S. Mendes, and E.K. Lenzi, 2001: Q-exponential Distribution in Urban Agglomeration. *Physical Review E*, **65**:1-3.
Moomaw, R.L., and A.M. Shatter, 1996: Urbanization and Economic Development: A Bias toward Large Cities? *Journal of Urban Economics*, **40**:13-37.
Nelson, C.H., and P.V. PreckelSource, 1989: The Conditional Beta Distribution as a Stochastic Production Function. *American Journal of Agricultural Economics*, **71**(2):370-378.
Nitsch, V., 2005: Zipf Zipped. *Journal of Urban Economics*, **57**:86-100.
Paltsev, S., J.M. Reilly, H.D. Jacoby, R.S. Eckaus, J. McFarland, M. Sarofim, M. Asadoorian, and M. Babiker, 2005: The MIT Emissions Prediction and Policy Analysis (EPPA) Model: Version 4. *Report 125*. MIT Joint Program on the Science and Policy of Global Change: Cambridge, MA. Available at: http://globalchange.mit.edu/files/document/MITJPSPGC_Rpt125.pdf.
Parr, J.B., 1970: Models of City Size in an Urban System. *Papers in Regional Science*, **25**(1):221-253.
Petrakos, G., P. Mardakis, and H. Caravel, 2000: Recent Developments in the Greek System of Urban Centres. *Environment and Planning B*, **27**:169-181.
Richardson, H.W., 1972: Optimality in City Size, Systems of Cities, and Urban Policy: A Skeptic's View. *Urban Studies*, **9**(1):29-48.
Rosen, K.T., and M. Resnick, 1980: The Size Distribution of Cities: An Examination of the Pareto Law and Primacy. *Journal of Urban Economics*, **8**:165-186.

Socioeconomic Data and Applications Center (SEDAC), 2005: *Gridded Population of the World, version 3, and the Global Rural-Urban Mapping Project*. [cited July 1, 2010]. Available at: http://sedac.ciesin.columbia.edu/gpw/.

Song, S., and K.H. Zhang, 2002: Urbanisation and City Size Distribution in China. *Urban Studies*, **39**(12):2317-2327.

Soo, K.T., 2007: Zipf's Law and Urban Growth in Malaysia. *Urban Studies*, **44**(1):1-14.

Suarez-Villa, L., 1988: Metropolitan Evolution, Sectoral Economic Change, and the City Size Distribution. *Urban Studies*, **25**(1):1-20.

Suh, S.H., 1991: The Optimal Size Distribution of Cities. *Journal of Urban Economics*, **30**(2):181-191.

United Nations, 2009: *World Urbanization Prospects: The 2009 Revision*. [cited September 30, 2011]. Available at: http://esa.un.org/unpd/wup/index.htm.

United Nations, 2010: *World Population Prospects: The 2010 Revision*. [cited November 1, 2011]. Available at: http://esa.un.org/unpd/wpp/Excel-Data/population.htm.

Venables, A.J., 2005: Spatial Disparities in Developing Countries: Cities, Regions, and International Trade. *Journal of Economic Geography*, **5**:3-21.

World Bank, 2011: *World Development Indicators Online*. [cited May 21, 2011]. Available at: http://publications.worldbank.org/WDI/.

Yezer, A.M.J., and R.S. Goldfarb, 1978: An Indirect Test of Efficient City Sizes. *Journal of Urban Economics*, **5**:46-65.

Zheng, X.-P., 2007: Measurement of Optimal City Sizes in Japan: A Surplus Function Approach. *Urban Studies*, **44**(5/6):939-951.

Zipf, G.K., 1949: *Human Behaviour and the Principle of Least Effort*. Addison-Wesley: Reading, MA.

FOR THE COMPLETE LIST OF JOINT PROGRAM REPORTS:
http://globalchange.mit.edu/pubs/all-reports.php

153. Climate Change, Mortality, and Adaptation: ***Evidence from Annual Fluctuations in Weather in the U.S.*** *Deschênes & Greenstone* August 2007
154. Modeling the Prospects for Hydrogen Powered Transportation Through 2100 *Sandoval et al.* February 2008
155. Potential Land Use Implications of a Global Biofuels Industry *Gurgel et al.* March 2008
156. Estimating the Economic Cost of Sea-Level Rise *Sugiyama et al.* April 2008
157. Constraining Climate Model Parameters from Observed 20th Century Changes *Forest et al.* April 2008
158. Analysis of the Coal Sector under Carbon Constraints *McFarland et al.* April 2008
159. Impact of Sulfur and Carbonaceous Emissions from International Shipping on Aerosol Distributions and Direct Radiative Forcing *Wang & Kim* April 2008
160. Analysis of U.S. Greenhouse Gas Tax Proposals *Metcalf et al.* April 2008
161. A Forward Looking Version of the MIT Emissions Prediction and Policy Analysis (EPPA) Model *Babiker et al.* May 2008
162. The European Carbon Market in Action: ***Lessons from the first trading period*** Interim Report *Convery, Ellerman, & de Perthuis* June 2008
163. The Influence on Climate Change of Differing Scenarios for Future Development Analyzed Using the MIT Integrated Global System Model *Prinn et al.* September 2008
164. Marginal Abatement Costs and Marginal Welfare Costs for Greenhouse Gas Emissions Reductions: ***Results from the EPPA Model*** *Holak et al.* November 2008
165. Uncertainty in Greenhouse Emissions and Costs of Atmospheric Stabilization *Webster et al.* November 2008
166. Sensitivity of Climate Change Projections to Uncertainties in the Estimates of Observed Changes in Deep-Ocean Heat Content *Sokolov et al.* November 2008
167. Sharing the Burden of GHG Reductions *Jacoby et al.* November 2008
168. Unintended Environmental Consequences of a Global Biofuels Program *Melillo et al.* January 2009
169. Probabilistic Forecast for 21st Century Climate Based on Uncertainties in Emissions (without Policy) and Climate Parameters *Sokolov et al.* January 2009
170. The EU's Emissions Trading Scheme: ***A Proto-type Global System?*** *Ellerman* February 2009
171. Designing a U.S. Market for CO_2 *Parsons et al.* February 2009
172. Prospects for Plug-in Hybrid Electric Vehicles in the United States & Japan: ***A General Equilibrium Analysis*** *Karplus et al.* April 2009
173. The Cost of Climate Policy in the United States *Paltsev et al.* April 2009
174. A Semi-Empirical Representation of the Temporal Variation of Total Greenhouse Gas Levels Expressed as Equivalent Levels of Carbon Dioxide *Huang et al.* June 2009
175. Potential Climatic Impacts and Reliability of Very Large Scale Wind Farms *Wang & Prinn* June 2009
176. Biofuels, Climate Policy and the European Vehicle Fleet *Gitiaux et al.* August 2009
177. Global Health and Economic Impacts of Future Ozone Pollution *Selin et al.* August 2009
178. Measuring Welfare Loss Caused by Air Pollution in Europe: ***A CGE Analysis*** *Nam et al.* August 2009
179. Assessing Evapotranspiration Estimates from the Global Soil Wetness Project Phase 2 (GSWP-2) Simulations *Schlosser and Gao* September 2009
180. Analysis of Climate Policy Targets under Uncertainty *Webster et al.* September 2009
181. Development of a Fast and Detailed Model of Urban-Scale Chemical and Physical Processing *Cohen & Prinn* October 2009
182. Distributional Impacts of a U.S. Greenhouse Gas Policy: ***A General Equilibrium Analysis of Carbon Pricing*** *Rausch et al.* November 2009
183. Canada's Bitumen Industry Under CO_2 Constraints *Chan et al.* January 2010
184. Will Border Carbon Adjustments Work? *Winchester et al.* February 2010
185. Distributional Implications of Alternative U.S. Greenhouse Gas Control Measures *Rausch et al.* June 2010
186. The Future of U.S. Natural Gas Production, Use, and Trade *Paltsev et al.* June 2010
187. Combining a Renewable Portfolio Standard with a Cap-and-Trade Policy: ***A General Equilibrium Analysis*** *Morris et al.* July 2010
188. On the Correlation between Forcing and Climate Sensitivity *Sokolov* August 2010
189. Modeling the Global Water Resource System in an Integrated Assessment Modeling Framework: IGSM-WRS *Strzepek et al.* September 2010
190. Climatology and Trends in the Forcing of the Stratospheric Zonal-Mean Flow *Monier and Weare* January 2011
191. Climatology and Trends in the Forcing of the Stratospheric Ozone Transport *Monier and Weare* January 2011
192. The Impact of Border Carbon Adjustments under Alternative Producer Responses *Winchester* February 2011
193. What to Expect from Sectoral Trading: ***A U.S.-China Example*** *Gavard et al.* February 2011
194. General Equilibrium, Electricity Generation Technologies and the Cost of Carbon Abatement *Lanz and Rausch* February 2011

195. A Method for Calculating Reference Evapotranspiration on Daily Time Scales *Farmer et al.* February 2011
196. Health Damages from Air Pollution in China *Matus et al.* March 2011
197. The Prospects for Coal-to-Liquid Conversion: ***A General Equilibrium Analysis*** *Chen et al.* May 2011
198. The Impact of Climate Policy on U.S. Aviation *Winchester et al.* May 2011
199. Future Yield Growth: ***What Evidence from Historical Data*** *Gitiaux et al.* May 2011
200. A Strategy for a Global Observing System for Verification of National Greenhouse Gas Emissions *Prinn et al.* June 2011
201. Russia's Natural Gas Export Potential up to 2050 *Paltsev* July 2011
202. Distributional Impacts of Carbon Pricing: ***A General Equilibrium Approach with Micro-Data for Households*** *Rausch et al.* July 2011
203. Global Aerosol Health Impacts: ***Quantifying Uncertainties*** *Selin et al.* August 201
204. Implementation of a Cloud Radiative Adjustment Method to Change the Climate Sensitivity of CAM3 *Sokolov and Monier* September 2011
205. Quantifying the Likelihood of Regional Climate Change: ***A Hybridized Approach*** *Schlosser et al.* Oct 2011
206. Process Modeling of Global Soil Nitrous Oxide Emissions *Saikawa et al.* October 2011
207. The Influence of Shale Gas on U.S. Energy and Environmental Policy *Jacoby et al.* November 2011
208. Influence of Air Quality Model Resolution on Uncertainty Associated with Health Impacts *Thompson and Selin* December 2011
209. Characterization of Wind Power Resource in the United States and its Intermittency *Gunturu and Schlosser* December 2011
210. Potential Direct and Indirect Effects of Global Cellulosic Biofuel Production on Greenhouse Gas Fluxes from Future Land-use Change *Kicklighter et al.* March 2012
211. Emissions Pricing to Stabilize Global Climate *Bosetti et al.* March 2012
212. Effects of Nitrogen Limitation on Hydrological Processes in CLM4-CN *Lee & Felzer* March 2012
213. City-Size Distribution as a Function of Socio-economic Conditions: ***An Eclectic Approach to Down-scaling Global Population*** *Nam & Reilly* March 2012