The Spatio-Temporal Regression Approach for Small-Area Population Forecasting

Guangqing Chi

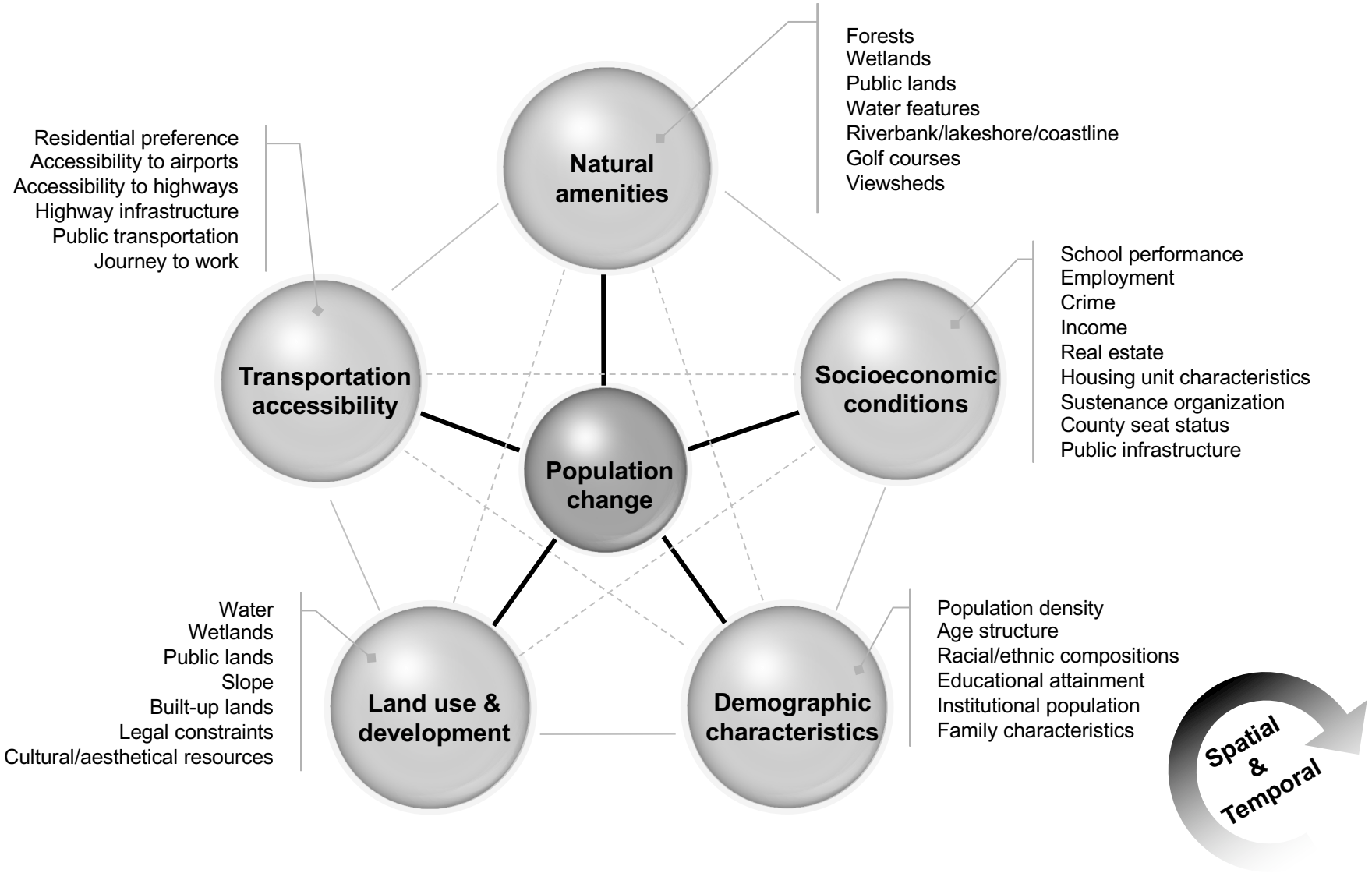
Department of Agricultural Economics, Sociology, and Education,
Population Research Institute, and Social Science Research Institute
Pennsylvania State University

The 26th Annual Population Research Institute Sponsored Demography Graduate Student Methodology Workshop, Penn State, May 14, 2019

Small-Area Population Forecasting: Fundamental Methods

- Extrapolation methods
- Cohort component methods
- Ratio methods
- Housing unit methods

(Knowledge-based) Regression Approach for Population Forecasting



The Spatio-Temporal Regression Approach for Population Forecasting

- Chi, Guangqing and Jun Zhu. 2019. *Spatial Regression Models for the Social Sciences*. SAGE Publications: Thousand Oaks, CA. Chapter 7.
- Chi, Guangqing and Donghui Wang. 2017. "Small-Area Population Forecasting: A Geographically Weighted Regression Approach." In: *The Frontiers of Applied Demography*, pp. 449–471, edited by David Swanson. New York: Springer.
- Chi, Guangqing and Paul R. Voss. 2011. "Small-Area Population Forecasting: Borrowing Strength across Space and Time." *Population, Space and Place* 17(5): 505–520.
- Chi, Guangqing. 2009. "Can Knowledge Improve Population Forecasts at Subcounty Levels?" *Demography* 46(2): 405–427.

A Standard Regression Forecasting Model

In step one, it establishes a relation between the response variable and its explanatory variables measured at an earlier time point, because forecasting is to project the future with current and past information.

$$Y_t = \alpha Y_{t-1} + X_{t-1}\beta + \varepsilon_t$$

where

- Y_t is an n by 1 vector of response variables at time t,
- Y_{t-1} is an n by 1 vector of response variables at time t–1,
- α is a scalar coefficient for Y_{t-1} at time t–1,
- X_{t-1} is an n by p design matrix of explanatory variables at time t–1,
- β is a p by 1 vector of regression coefficients for X_{t-1} at time t–1, and
- ε_t is an n by 1 vector of error terms.

A Standard Regression Forecasting Model

In step two, we use the estimated coefficients, the explanatory variables at time t, and the response variable at time t to forecast the response variable at time t+1.

$$\hat{Y}_{t+1} = \hat{\alpha} Y_t + X_t \hat{\beta}$$

where

$\hat{Y}_{t+1}$ is an n by 1 vector of estimated response variables at time $t+1$,
Y_t is an n by 1 vector of observed response variables at time t,
X_t is an n by p design matrix of observed explanatory variables at time t, and
$\hat{\alpha}$ and $\hat{\beta}$ are the estimates of the coefficients from fitting the model

Ordinary least squares (OLS) regression models

$$y = X\beta + \varepsilon$$

Where
Y is an *n* by 1 vector of *n* observations on the response variable,
X is an *n* by *p* design matrix with a vector of *n* ones in the first column for the intercept and *p-1* vectors (*n* by 1) of explanatory variables in the remaining columns,
β is a *p* by 1 vector of regression coefficients, and
ε is an *n* by 1 vector of *n* error terms that are independently and identically distributed as normal distribution with mean 0 and a constant variance.

Spatial Error Model

$$y = X\beta + u$$

$$u = \rho W u + \varepsilon$$

where

Y is an n by 1 vector of response variables,
X is an n by p design matrix of explanatory variables,
β is a p by 1 vector of regression coefficients,
u is an n by 1 vector of error terms,
ρ is a scalar spatial error parameter,
W is an n by n spatial weight matrix, and
ε is an n by 1 vector of error terms that are normally and independently but not necessarily identically distributed.

How to Fit the Spatio-Temporal Regression Model

$$Y_t = X_t\beta + u_t,\ u_t = \rho W u_t + \varepsilon_t$$

$$Y_t = X_t\beta_1 + \beta_2 Y_{t-1} + X_{t-1}\beta_3 + u_t,\ u_t = \rho W u_t + \varepsilon_t$$

$$Y_t = Y_{t-1}\beta_2 + \rho W Y_t - \rho\beta_2 W Y_{t-1} + X_t\beta_1 + X_{t-1}\beta_3 - W X_t \rho\beta_1 - W X_{t-1}\rho\beta_3 + \varepsilon_t$$

The Spatio-Temporal Regression Model

$$Y_t = X_t\beta + \rho WY_t - \rho WX_t\beta + \tau_1 Y_{t-1} + X_{t-1}\tau_2 + \tau_3 WY_{t-1} + WX_{t-1}\tau_4 + \varepsilon_t$$

- explanatory variables,
- a spatially lagged response variable,
- spatially lagged explanatory variables,
- a temporally lagged response variable,
- temporally lagged explanatory variables,
- a spatially and temporally lagged response variable,
- spatially and temporally lagged explanatory variables, and
- a random error term.

The Spatio-Temporal Regression Model

$$Y_t = X_t\beta + \rho WY_t - \rho WX_t\beta + \tau_1 Y_{t-1} + X_{t-1}\tau_2 + \tau_3 WY_{t-1} - WX_{t-1}\tau_4 + \varepsilon_t$$

The Spatial Regression Forecasting Model

$$Y_t = \rho WY_t + \tau_1 Y_{t-1} + X_{t-1}\tau_2 + \tau_3 WY_{t-1} - WX_{t-1}\tau_4 + \varepsilon_t$$

The Spatial Regression Forecasting Model

$$Y_t = \rho WY_t + \tau_1 Y_{t-1} + X_{t-1}\tau_2 + \tau_3 WY_{t-1} - WX_{t-1}\tau_4 + \varepsilon_t$$

$$Y_t = (I - \rho W)^{-1}(\tau_1 Y_{t-1} + X_{t-1}\tau_2 + \tau_3 WY_{t-1} - WX_{t-1}\tau_4 + \varepsilon_t)$$

- the temporally lagged response variable (Y_{t-1}),
- temporally lagged explanatory variables (X_{t-1}),
- temporally lagged weighted average of the response variable in the neighborhood (WY_{t-1}), and
- temporally lagged weighted average of the explanatory variables in the neighborhood (WX_{t-1}).

$$Y_{1990,2000} = \alpha Y_{1980,1990} + X_{1990}\beta + \lambda W Y_{1980,1990} + W X_{1990}\rho + \varepsilon$$

Table 7.4 Model fitting results of initial spatial regression

	Coefficient	S.E.
Previous growth	0.230***	0.029
Old	0.105	0.065
Unemployment	0.057	0.101
Airport	-0.089	0.116
Forest	0.152***	0.031
Land developability	0.030	0.023
Spatially lagged previous change	0.627***	0.050
Spatially lagged old	0.207	0.112
Spatially lagged unemployment	-0.171	0.168
Spatially lagged airport	0.179	0.144
Spatially lagged forest	-0.134***	0.038
Spatially lagged land developability	0.033	0.026
Measures of fit		
Log-likelihood	1005.240	
AIC	-1986.480	
BIC	-1920.290	
n	1,837	

Notes: * Significant at the $\alpha = 0.05$ level; ** Significant at the $\alpha = 0.01$ level; *** Significant at the $\alpha = 0.001$ level. AIC = Akaike's information criterion. BIC = Schwartz's Bayesian information criterion. S.E. = standard error.

Four Finalized Population Forecasting Models

Baseline: Extrapolation projection

$$G = \left[\frac{P_{90} - P_{80}}{10} + \frac{P_{90} - P_{70}}{20} + \frac{P_{90} - P_{60}}{30} \right] \Big/ 3$$

$$P_{2000} = P_{90} + 10 \times G$$

Model 1: Standard regression

$$Ln\left(\frac{P_{90}}{P_{80}}\right) = \left[Ln\left(\frac{P_{80}}{P_{70}}\right) + X_{80} \right] \beta + \varepsilon$$

$$\overbrace{Ln\left(\frac{P_{00}}{P_{90}}\right)} = \left[Ln\left(\frac{P_{90}}{P_{80}}\right) + X_{90} \right] \hat{\beta}$$

Model 2: partial spatio-temporal regression (incorporating spatial population effects)

$$Ln\left(\frac{P_{90}}{P_{80}}\right) = \left[Ln\left(\frac{P_{80}}{P_{70}}\right) + X_{80} \right] \beta + \lambda WLn\left(\frac{P_{80}}{P_{70}}\right)_{neighbor} + \varepsilon$$

$$\overbrace{Ln\left(\frac{P_{00}}{P_{90}}\right)} = \left[Ln\left(\frac{P_{90}}{P_{80}}\right) + X_{90} \right] \hat{\beta} + \hat{\lambda} WLn\left(\frac{P_{90}}{P_{80}}\right)_{neighbor}$$

Model 3: full spatio-temporal regression (incorporating spatial population effects and other neighbor characteristics)

$$Ln\left(\frac{P_{90}}{P_{80}}\right) = \left[Ln\left(\frac{P_{80}}{P_{70}}\right) + X_{80} \right] \beta + \lambda_1 WLn\left(\frac{P_{80}}{P_{70}}\right)_{neighbor} + \lambda_2 W(X_{80})_{neighbor} + \varepsilon$$

$$\overbrace{Ln\left(\frac{P_{00}}{P_{90}}\right)} = \left[Ln\left(\frac{P_{90}}{P_{80}}\right) + X_{90} \right] \hat{\beta} + \hat{\lambda}_1 WLn\left(\frac{P_{90}}{P_{80}}\right)_{neighbor} + \hat{\lambda}_2 W(X_{90})_{neighbor}$$

Estimations of three regression models

	Model 1 (Standard regression)		Model 2 (Regression with neighbor growth)		Model 3 (Regression with neighbor growth and neighbor characteristics)	
	Coef.	p-value	Coef.	p-value	Coef.	p-value
Growth rate in previous decade	0.135	0.000	0.105	0.000	0.092	0.000
Population density	0.006	0.000	0.007	0.000	0.004	0.015
Young	–0.356	0.000	–0.392	0.000	–0.705	0.000
House value	0.001	0.000	0.001	0.000	0.000	0.000
New housing	–0.067	0.000	–0.039	0.022	–0.060	0.001
Growth rate in previous decade (neighbor average)	/	/	0.201	0.000	0.194	0.000
Young (neighbor average)	/	/	/	/	0.415	0.000
Measures of fit						
Adjusted R^2	0.1791		0.2029		0.2200	
Log likelihood	1446.26		1474.28		1485.77	
AIC	–2882.53		–2936.56		–2957.55	
Remaining spatial dependence						
Robust LM (error)	1.533	0.22	11.107	0.00	0.069	0.79
Robust LM (lag)	17.766	0.00	2.151	0.14	2.696	0.10

Evaluating population projections

	Baseline projection (Extrapolation projection)	Model 1	Model 2	Model 3
MALPE Measure of bias	–3.65%	–3.38%	–3.39%	–3.34%
MAPE Measure of precision	9.63%	10.79%	10.78%	10.78%
RMSPE	13.56%	15.06%	15.03%	15.10%

Evaluating population projections by population size in 2000

Population size (Number of MCDs)	250 and less (118)		251 — 2,000 (1,310)		2,001 — 20,000 (372)		20,001 and more (37)	
	MALPE	MAPE	MALPE	MAPE	MALPE	MAPE	MALPE	MAPE
Baseline projection	3.47%	16.16%	-3.99%	9.41%	-5.03%	8.84%	-0.81%	4.61%
Model 1	3.42%	15.46%	-3.51%	10.30%	-5.33%	11.51%	-0.77%	6.16%
Model 2	3.46%	15.39%	-3.50%	10.30%	-5.43%	11.47%	-1.04%	6.25%
Model 3	4.16%	15.56%	-3.50%	10.28%	-5.38%	11.47%	-1.06%	6.08%

Evaluating population projections by population growth rate from 1990 to 2000

Population growth rate (Number of MCDs)	-10% and less (97)		-10% — -5.01% (105)		-5% — -0.01% (184)		0% (8)	
	MALPE	MAPE	MALPE	MAPE	MALPE	MAPE	MALPE	MAPE
Baseline projection	23.25%	23.81%	8.95%	9.47%	4.52%	5.87%	-1.48%	4.20%
Model 1	26.38%	26.38%	11.80%	11.87%	7.06%	7.73%	2.73%	3.57%
Model 2	26.31%	26.31%	11.79%	11.86%	7.03%	7.65%	3.21%	3.47%
Model 3	26.53%	26.53%	11.47%	11.54%	6.98%	7.69%	3.44%	4.77%

Population growth rate (Number of MCDs)	0.01% — 4.99% (283)		5% — 9.99% (299)		10% and more (861)	
	MALPE	MAPE	MALPE	MAPE	MALPE	MAPE
Baseline projection	0.43%	4.50%	-3.26%	5.22%	-11.47%	12.13%
Model 1	1.66%	4.46%	-2.28%	4.96%	-12.91%	13.72%
Model 2	1.71%	4.42%	-2.32%	4.94%	-12.93%	13.75%
Model 3	1.66%	4.67%	-2.26%	4.93%	-12.81%	13.66%

Can Knowledge Improve Forecasts?

- Things just didn't turn out as we hypothesized (and hoped) they would
- Our fancy spatio-temporal regression model outperformed simple regression in the *estimation* stage of the analysis
  - but who cares?
- In the *forecasting* stage, our spatio-temporal regression model does not outperform the extrapolation projection.

Why does not the theoretically grounded spatio-temporal regression approach outperform a simple atheoretical extrapolation projection?

- Temporal instability in the relationships between the estimation period and the forecasting period might undermine the promise of the regression-based approaches.

POPULATION AND DEVELOPMENT REVIEW

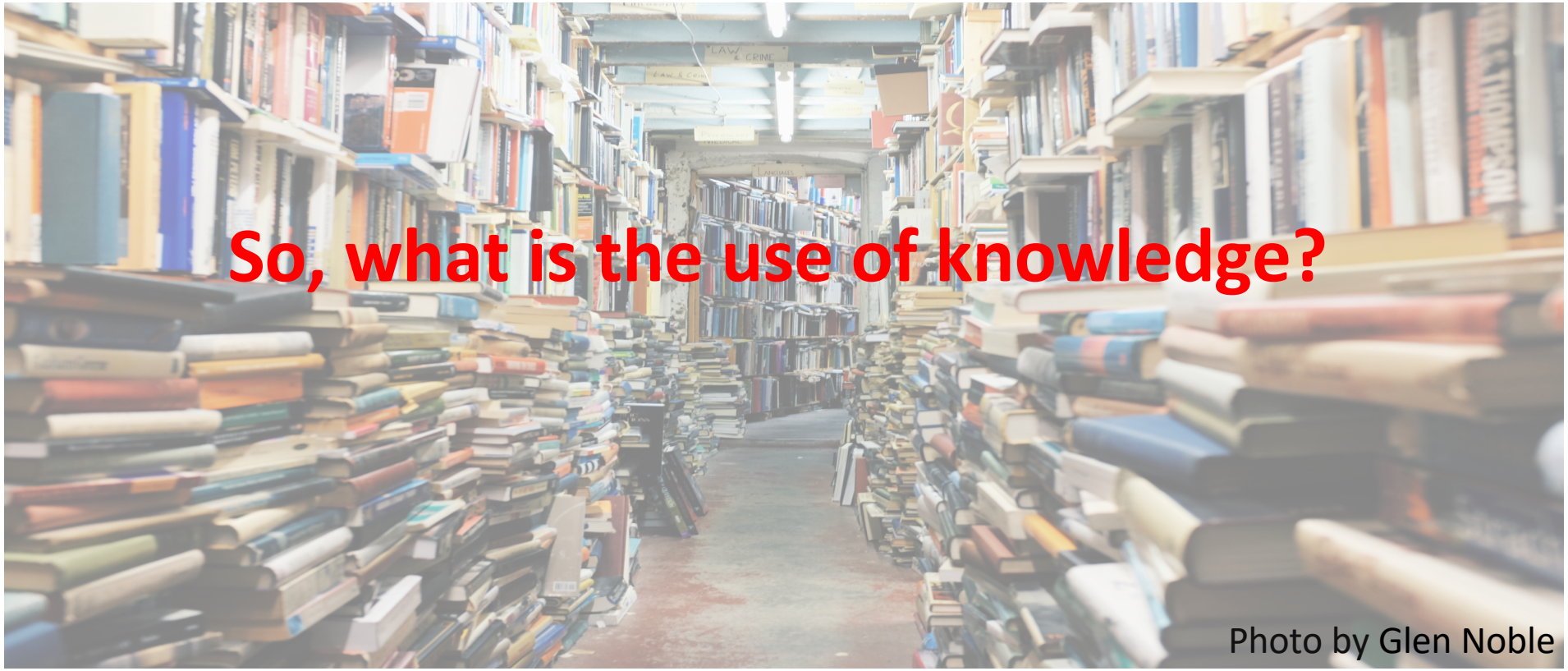
Can Knowledge Improve Forecasts?
Author(s): Nathan Keyfitz
Source: *Population and Development Review*, Vol. 8, No. 4 (Dec., 1982), pp. 729-751

"Forecasts of weather and earthquakes, where the next few hours are the subject of interest, and of unemployment, where the next year or two is what counts, are difficult enough. Population forecasts, where one peers a generation or two ahead, are even more difficult." (Keyfitz, 1982:746).

POPULATION AND DEVELOPMENT REVIEW

The Limits of Population Forecasting
Author(s): Nathan Keyfitz
Source: *Population and Development Review*, Vol. 7, No. 4 (Dec., 1981), pp. 579-593

"However interdisciplinary we become, there are some clear limits to knowledge of the interrelationships of the variables whose combined operation will bring about the future population." (Keyfitz, 1981:579).



Using the Spatial Regression Approach to Evaluate the Accuracy of Population Forecasting

Chi, Guangqing and Donghui Wang. 2018. "Population Projection Accuracy: The Impacts of Sociodemographics, Accessibility, Land Use, and Neighbor Characteristics." *Population, Space and Place* 24(5): e2129.

Prior Research

1. Population projection
   - Lots of research and methods (e.g., extrapolation, cohort component methods, spatial Bayesian, regression, etc)
2. Projection accuracy evaluation
   - Population size
   - Population growth rate
   - Region

What factors affect projection accuracy?

1. Associated factors (driving factors of population change)
   - Demographic characteristics
   - Socioeconomic conditions
   - Built environment and transportation accessibility
   - Natural environment
   - Land use and development
2. Neighbor characteristics
   - Spatial spillover effects
3. Spatial variations of the effects

Research Questions, Data, and Methods

1. Data
   - Decennial censuses 1970-2010
   - American Community Survey 2008-2012 estimates
   - Uniform Crime Reports of FBI, 2000
   - Land developability
2. Research questions and methods
   1) What characteristics do the predictable counties have? What characteristics do their neighbor counties have?
      - Standard regression, spatial lag models, and spatial error models
   2) Are the effects of the characteristics universal across the continental US? If not, how do they vary?
      - Geographically weighted regression (GWR)

Population Projection

- Baseline projection: a 40-year extrapolation projection based on the arithmetic change of populations in 1970,1980, 1990 and 2000 to project the population in 2010

- Percentage Error (PE) =

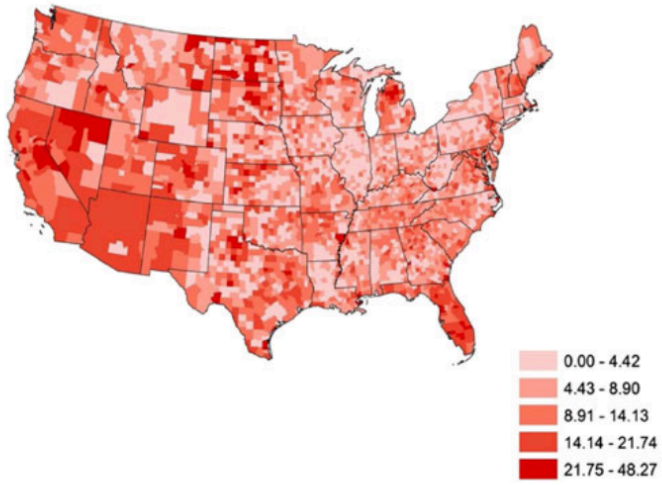
$$\frac{\text{Projected population}_{2010} - \text{Actual population}_{2010}}{\text{Actual population}_{2010}} \times 100\%$$

- Absolute Percentage Error (APE) =

$$\left| \frac{\text{Projected population}_{2010} - \text{Actual population}_{2010}}{\text{Actual population}_{2010}} \right| \times 100\%$$

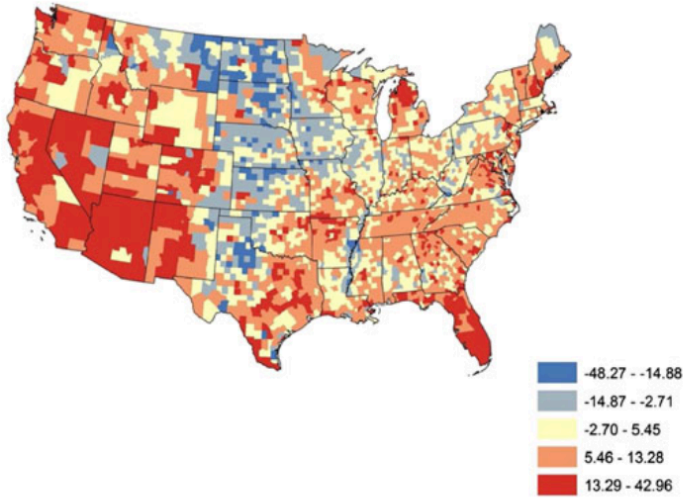
Distribution of

(a) APEs

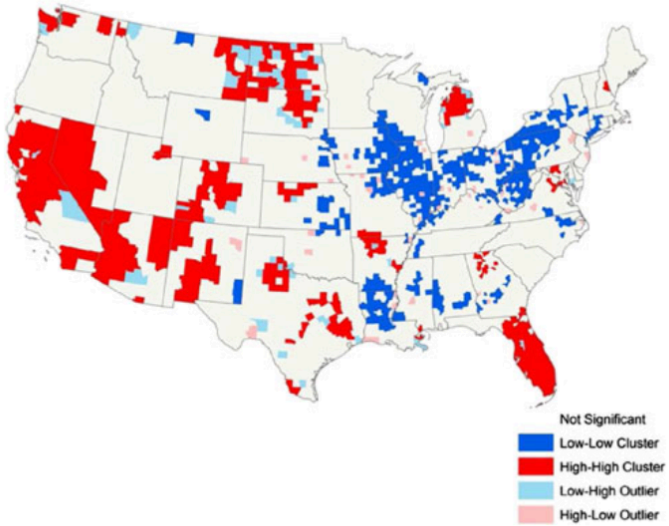


(a)

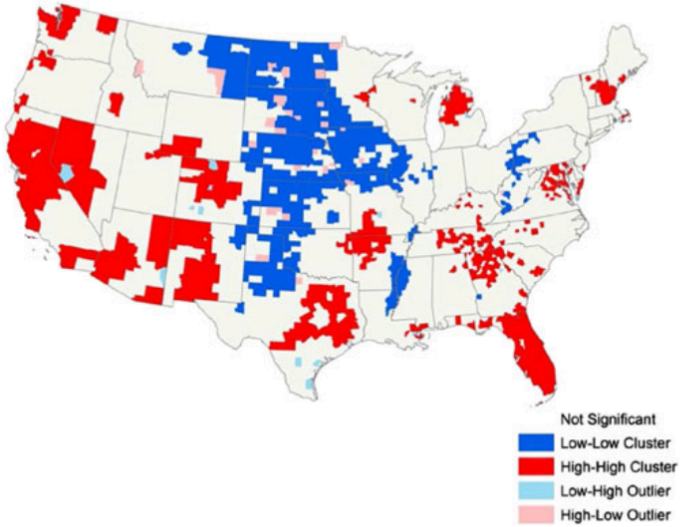
(b) PEs



(b)



(a)



(b)

What characteristics do the predictable counties have?

Demographic characteristics

Population size
Population growth rate
Population density
% young population (Age 12-18)
% old population (Age 65+)
% Black population
% Hispanic population
% workers in retail industry
% workers in agricultural industry

Socioeconomic conditions

The employment rate
% college population
% population (Age 25+) who finished high school
% population (Age 25+) with Bachelor's degree
Total crime rate
Violence crime rate

Transportation accessibility

% workers using public transportation to work
Journey to work (% workers traveling 30 minutes and less to work)
The inverse distance from the centroid of a county to its nearest major airport

Land use and development

County characteristics

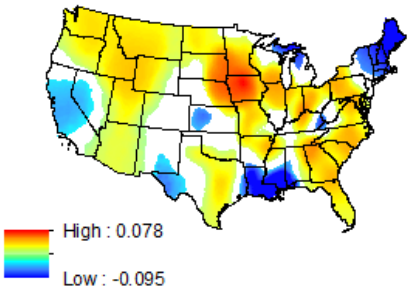
	Model 1 (APE)	Model 2 (PE)
Population growth rate	+	–
Population size	–	+
Young		+
Old		–
Black	+	
College	+	+
High school	+	
Bachelor's degree	+	+
Employed	–	
Agriculture	+	–
Retail		+
Income		–
Public transportation		–
Access to work	–	–
Land developability	–	
Airport accessibility	+	

Characteristics of neighboring counties

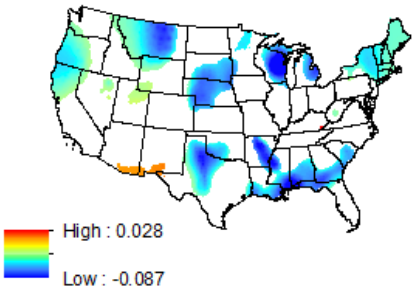
	Model 1 (APE)	Model 2 (PE)
Population growth rate	–	+
Young		–
Old		+
Hispanic	+	
High school	–	
Bachelor's degree		–
Employed		–
Agriculture	–	+
Income	–	
Access to work		+
Airport accessibility		+
Spatially lagged PE		+
Spatially lagged APE	+	

Local coefficients of the refined GWR model (dependent variable =APE)

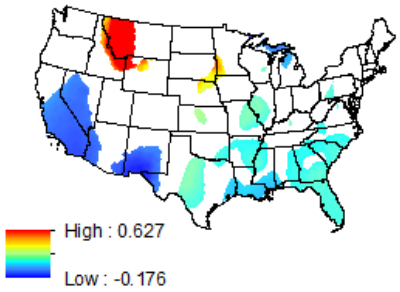
a. Population growth rate



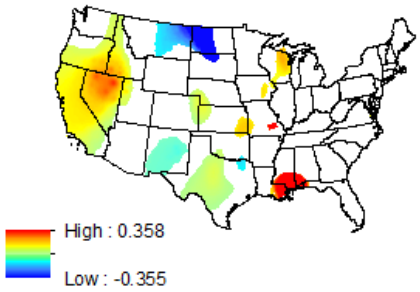
b. Population size



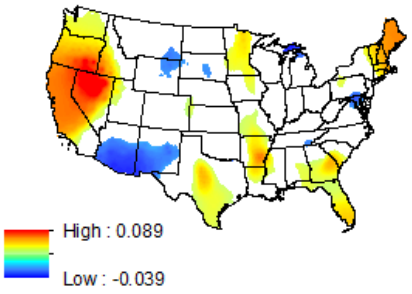
c. Black



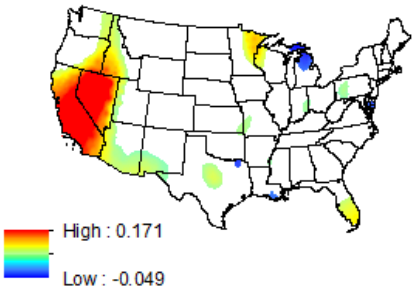
d. Hispanic



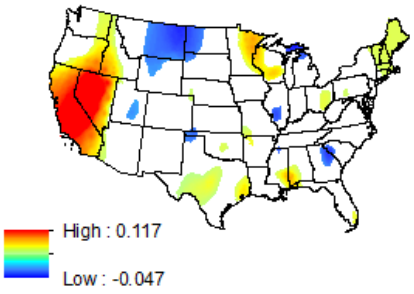
e. College



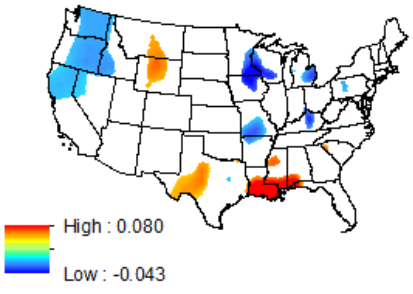
f. Highschool



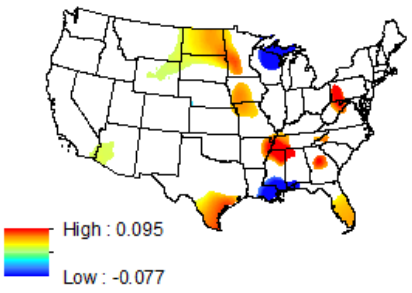
g. Bachelor's degree



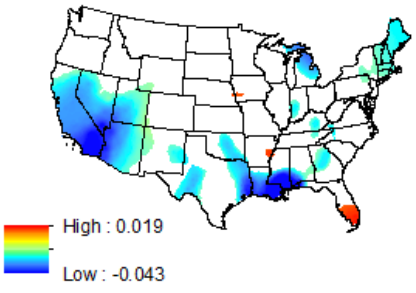
h. Employed



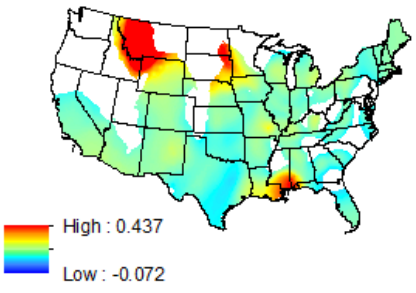
i. Agriculture



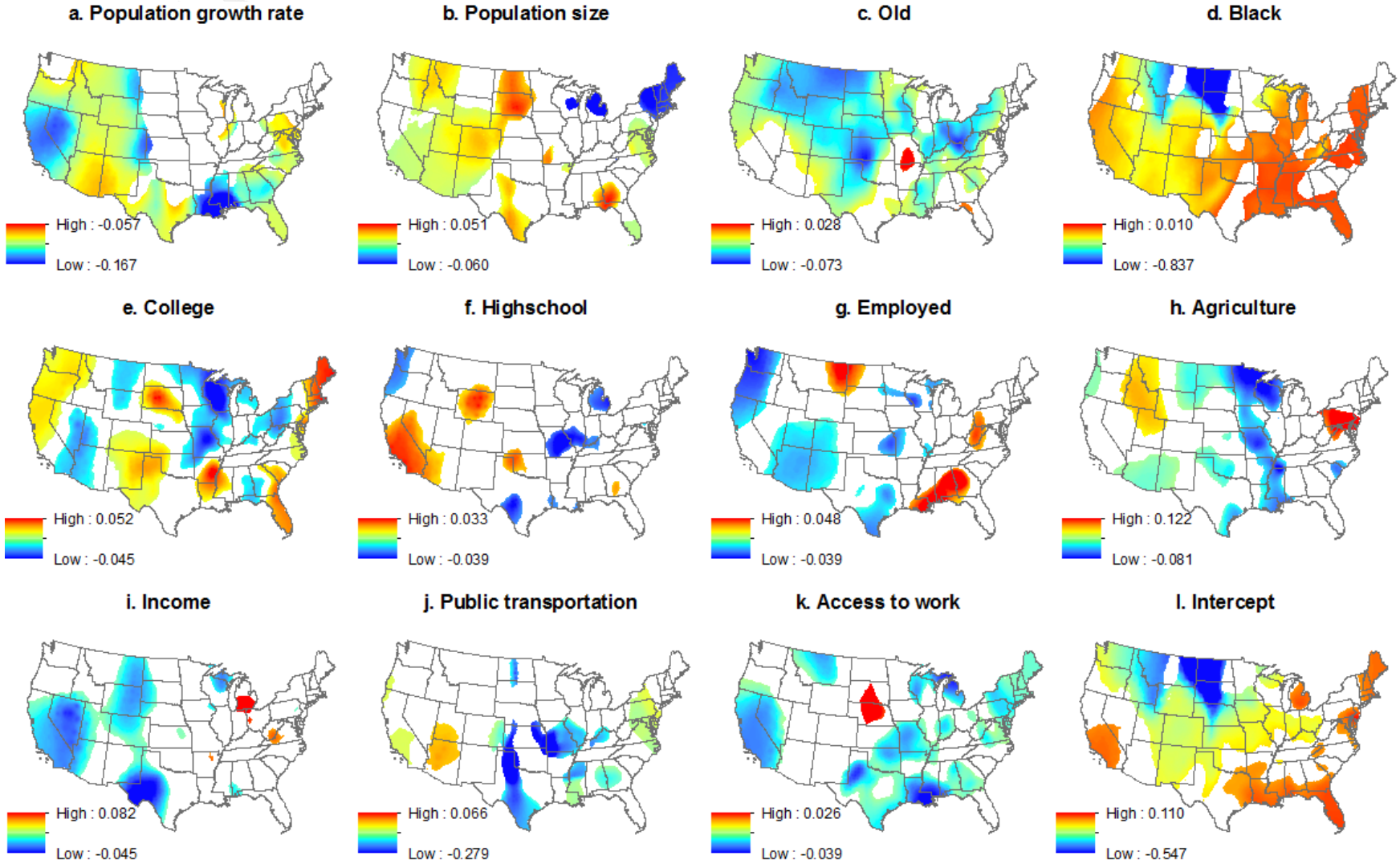
j. Access to work



k. Intercept

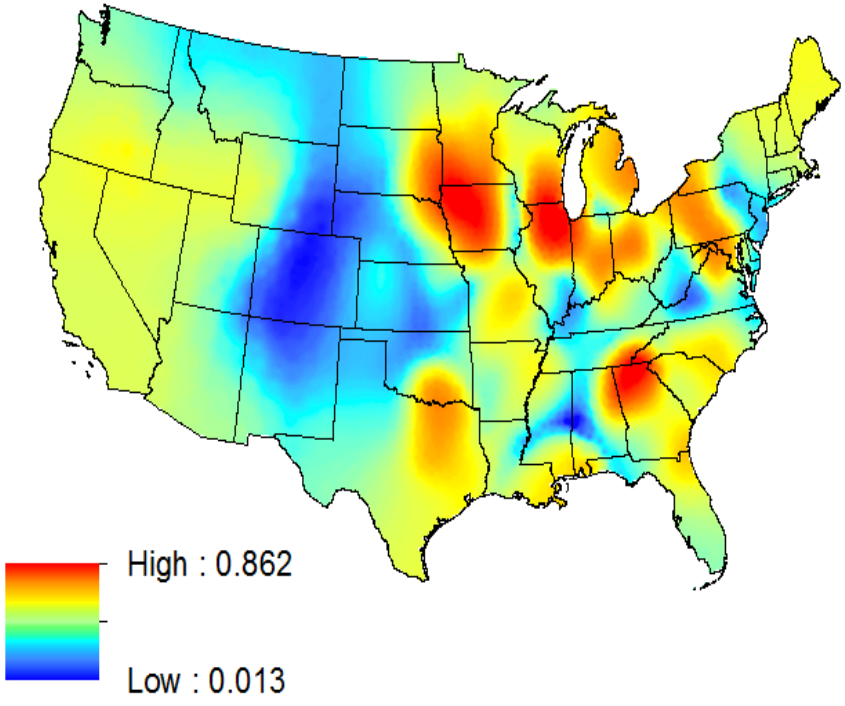


Local coefficients of the refined GWR model (dependent variable =PE)

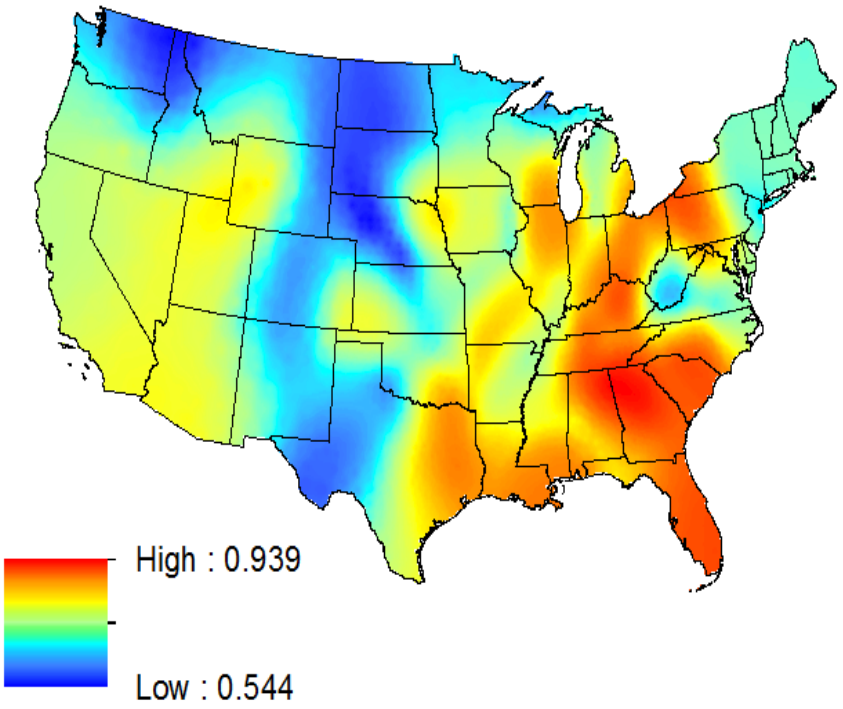


Local R^2 of the GWR models for (a) APE and (b) PE as a dependent variable

APE



PE



Conclusions

1. Who are predictable?
   - The counties whose populations are more predictable tend to be desirable places—places with abundant employment opportunities, decent public transportation infrastructure, easy accessibility to workplaces, and/or high land-development potential.
   - Their neighboring counties tend to have a well-educated population and a high income level.
2. Why are desirable places predictable?
   - They are attractive and often experience *stable* population growth (versus unpredictable changes), thus they are easier to predict.
3. However, this finding varies spatially, as some factors explain projection accuracy in different directions, magnitudes, or statistical significances across space.

Implications

- How population projections can be used *wisely* for urban and regional planning purposes.
  - The accuracy of population projections varies from one place to another.
  - To use population projections wisely, it is important to know how well a projection performs, where it performs better, what affects its accuracy, and how it interacts with neighboring counties.

Take-Home Messages

- Knowledge may not be able to help improve population projection accuracy.
- But knowledge can help with evaluating the projections and using them wisely.