
1st Report
Predicting city growth with machine learning
Simon Buechler, Anne Thompson, Dongxiao Niu, David Maroti

1 Strategic Background

2 Location Quotient

3 Machine Learning Tools

4 Results

5 Summary

City Growth Potential

Project is structured in a two-stage analysis:

- Identify which Chinese and US cities will grow in population and what drives this growth
- Analyze the link between the city's growth potential and commercial and residential real estate prices

- In the first stage we:
  - Investigate the link between city growth, industry growth, transportation infrastructure, human capital, entrepreneurship, and amenities
  - We use the most recent machine learning models to predict city growth and causally identify the main drivers

05.20

- Gather Data
- Specify methodology

06-08.20

- Monthly Webinar
- Start **1st** stage analysis
- Knowledge mapping on urban growth

09-11.20

- Start **2nd** stage analysis
- Present initial findings at international conferences

01.21

- Finish first draft of academic papers

05-08.21

- Final report and Presentation
- Submit two academic papers

City Growth Projection |
Defining the Location Quotient (LQ)

- The higher the LQ the greater the significance of this industry for the export base of the city

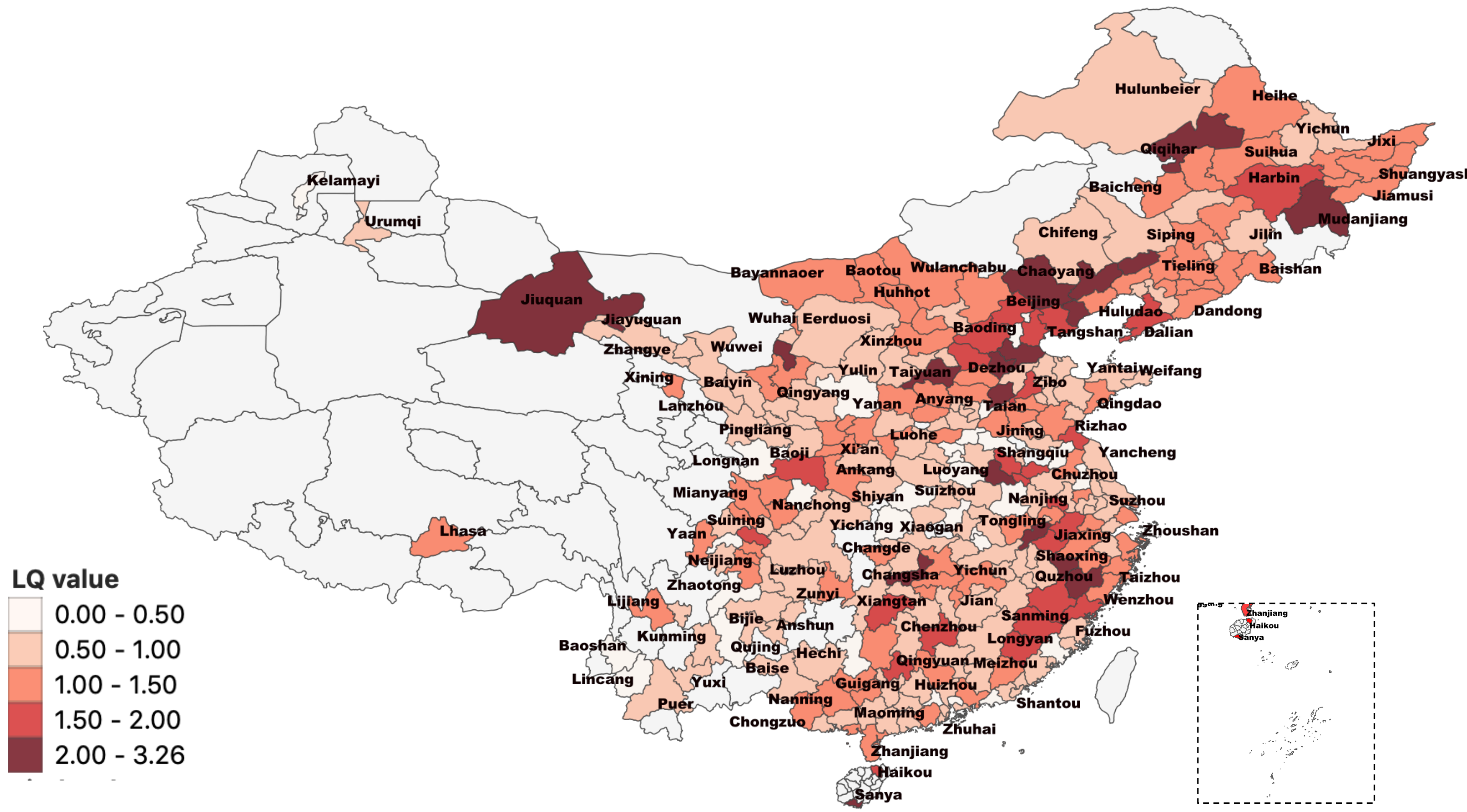
- Cities with high LQ's in growing industries are expected to grow the most

- LQ>1: city-level industry growth rate > national growth rate

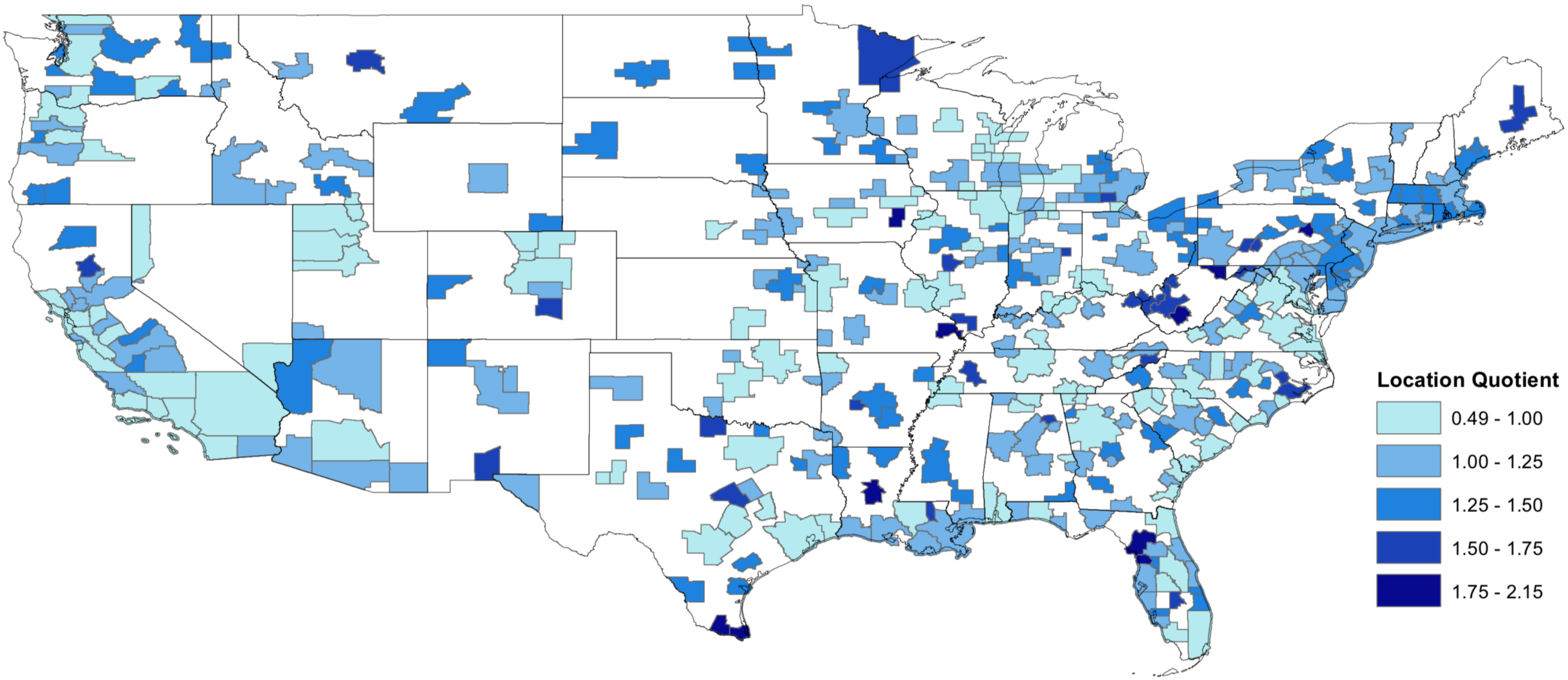
China LQ (1)| Information Communication, Computer Service & Software

China LQ (2)| Real Estate

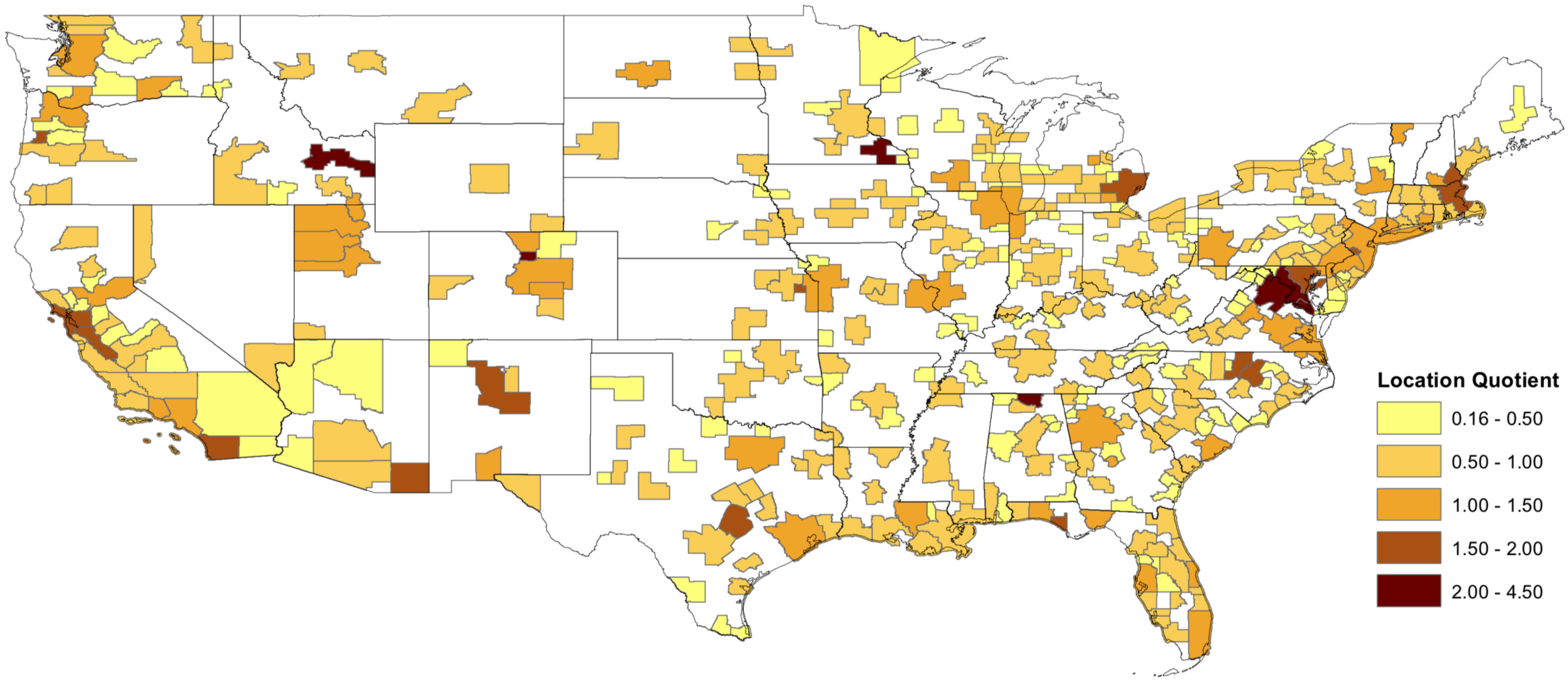
China LQ (3)| Finance & Insurance



US LQ (1)| Healthcare and Social Assistance



US LQ (2)| Professional, Scientific and Technical Services



City Growth Projection | Machine Learning Parameters (1)

- **Machine learning** constructs algorithms that can learn from the data

- **Big data** can come in two forms:
  - Wide (high-dimensional) data:
    - Many predictors (large p) and relatively small N
    - Typical method: Regularized regression
  - Tall or long data:
    - Many observations, but only few predictors
    - Typical method: Tree-based methods

✓ **Wide Data**: Many city growth drivers for a relatively small amount of cities

✓ **Regularized regression:** Method for selecting and fitting predictors that appear in a model

City Growth Projection |
Machine Learning Parameters (2)

- **Supervised Machine Learning:** You have an outcome Y and predictors X
  - Classical ML setting: independent observations
  - You fit the model Y that you want to predict using unseen data X0

- **Unsupervised Machine Learning:**
  - No pre-existing labels, undetected patterns
  - Dimension reduction: reduce the complexity of your data
  - Can be used to generate inputs (features) for supervised learning (e.g. Principal component regression)

✓ **Supervised Machine Learning**

- Focus on prediction
- Typical problems:
  - Netflix: predict user-rating of films
  - Predicting city growth
- Procedure: Algorithm is trained and validated using "unseen" data
- Strengths: Out-of-sample prediction, high-dimensional data, data-driven model selection

City Growth Projection | Regression Model

"Changes on levels" regression:

$$\Delta_{t+1,t} \log N_i = \lambda\beta_0 - \lambda \log N_{it} - \lambda\beta_I \log D_i + \epsilon_{it}$$

Key Components

- We do not know the true model. *Which regressors are important?*
- *How many regressors to include?*
    - Including too many regressors leads to overfitting: good in-sample fit (high R2), but bad out-of-sample prediction
    - Including too few regressors leads to omitted variable bias

Wide data adds complexity & makes model selection even more challenging

City Growth Projection | Estimation Methods

- **Regularized regression** removes some predictors from the model (i.e., forcing some coefficients to be zero) by choosing the penalization level lambda
- **Relevant** predictors can be chosen with cross-validation (CV)
- CV is a generalization where the data is iteratively split in training and validation sample
- **CV selects the lambda** (penalization) that minimizes an estimate of the out-of-sample prediction error

- **OLS**: include all regressors and minimizes mean square errors
- **LASSO:** Least Absolute Shrinkage and Selection Operator - (Tibshirani 1996) with penalty level (lambda) selected by information criteria EBIC (Chen and Chen 2008)

Parameters of analysis | Variables Embedded into China's Growth Model

Transportation Infrastructure

- Kilometers of bus lanes per capita
- Kilometers of highway per-capita

Industry Growth

- Growth in 20 different industries

Entrepreneurship

- Number of firms per-capita

Human Capital

- Number of colleges
- Share of highly-educated population
- Share of median income
- Patent

Amenities

- Days of snow
- Number of hospitals

Output

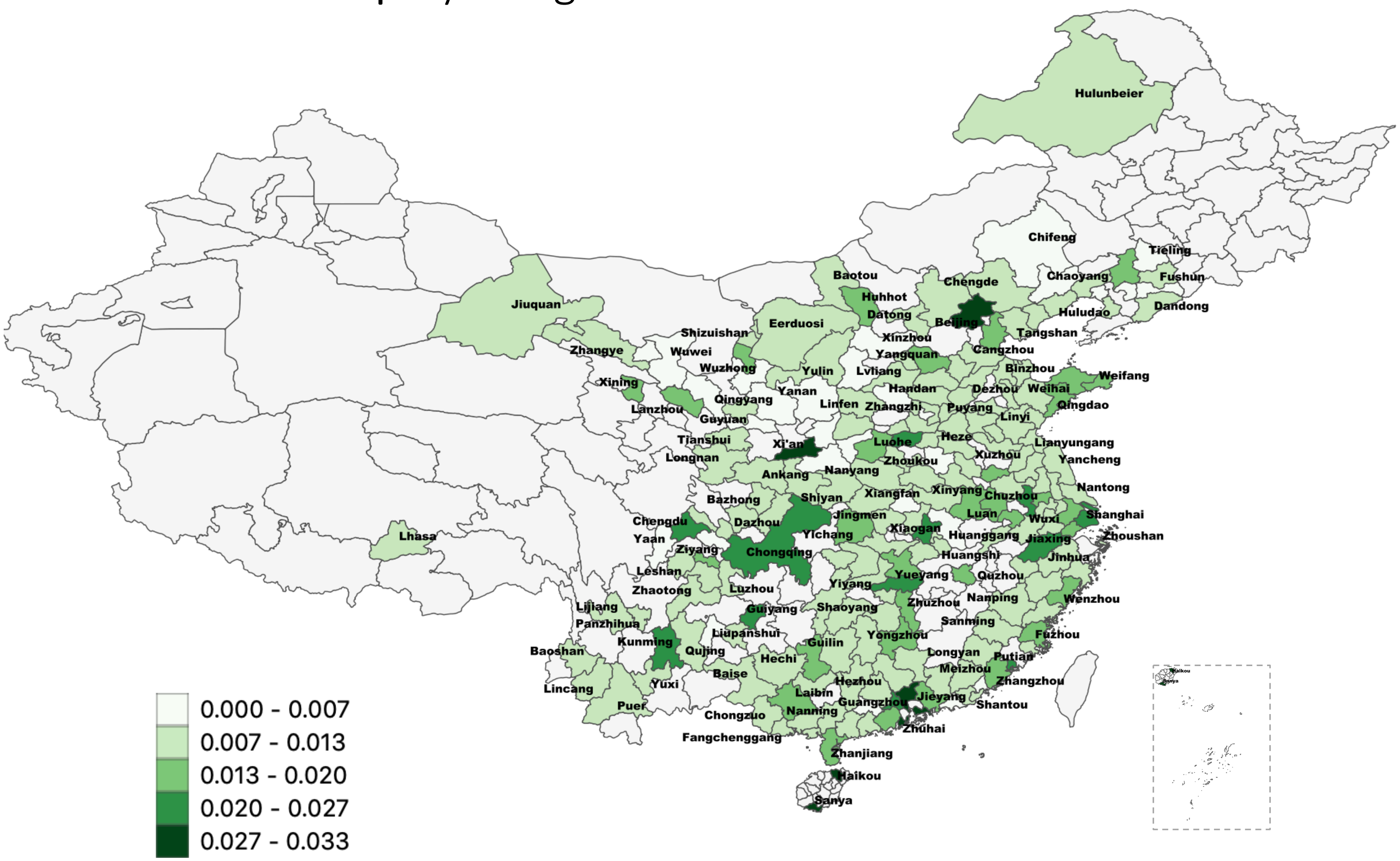
- GDP per capita
- Disposable income
- Government Income

Parameters of analysis | Variables Embedded into China's Growth Model

	Factors	*2 years*	*5 years*	*10 years*
Output	GDP per Capita	+	+	+
	Disposable Income			-
Human Capital	Patents	+		+
	College		+	+
	Share of Median Income			-
Employment	Information & Computer	+		
	Household Service		+	
	R&D	+	+	
	Transportation, Storage and post			+
	Real Estate	+	+	+
	Accommodation and Catering			-
	Leasing and Business Services			-
	Manufacturing			-
	Water, Conservancy, Environment			-

	Factors	*2 years*	*5 years*	*10 years*
	Roads			+
Amenities	Taxi		+	+
	Days of Fog			+
	Days of Storm			+
	Precipitation			+
	Max Temperature		+	
	Temperature		+	
	Hospitals			+
	Hospital Beds			+
	Bus		+	-
	Doctors			-
	Days of Frost	-		
	Days of Snow		-	

Prediction Results | 2 years growth in China



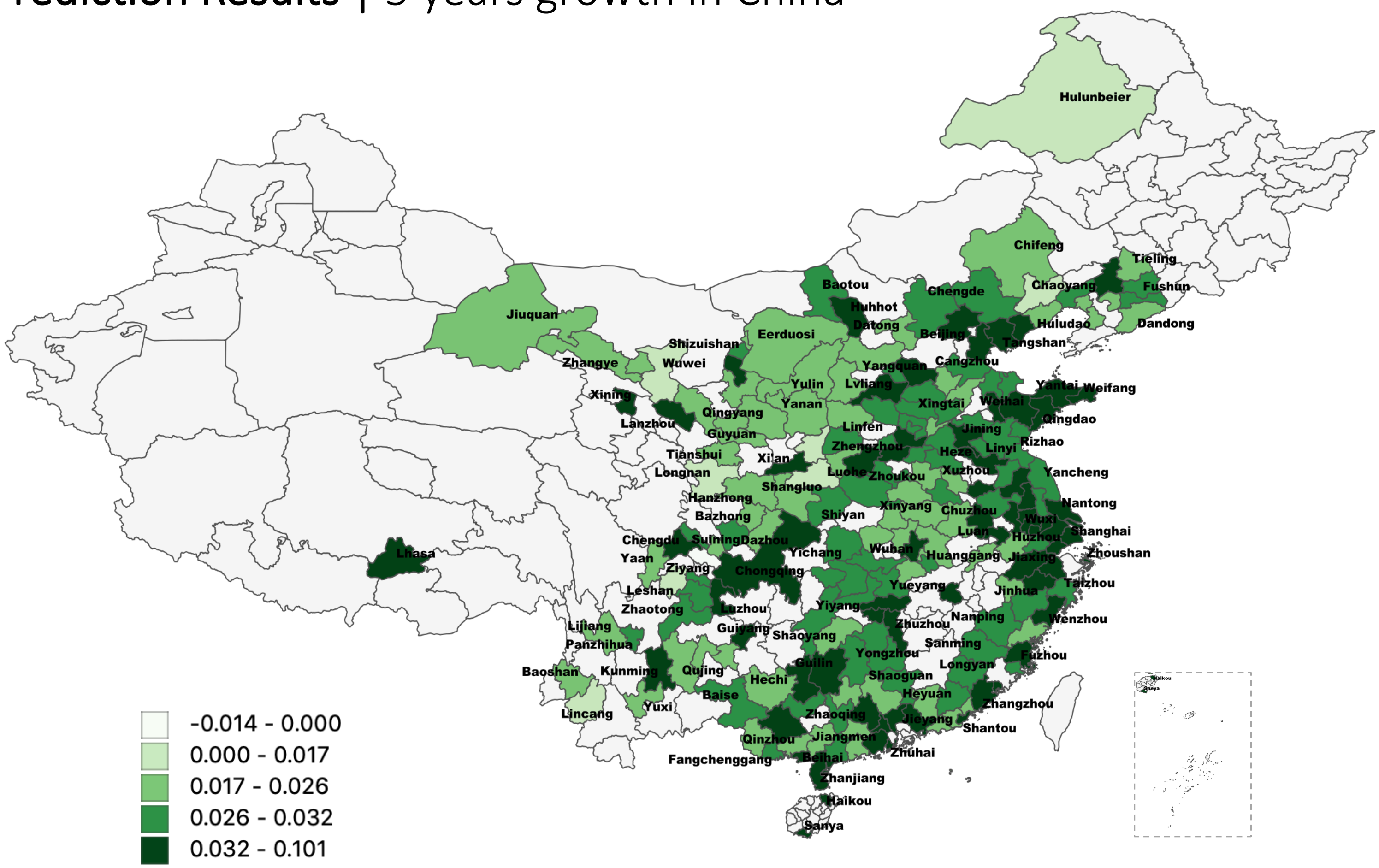
Top 5

- Xi'an 西安
- Guangzhou 广州
- Haikou 海口
- Shenzhen 深圳
- Beijing 北京

Bottom 5

- Lvliang 吕梁
- Xinzhou 忻州
- Jinchang 金昌
- Chaoyang 朝阳
- Weinan 渭南

Prediction Results | 5 years growth in China



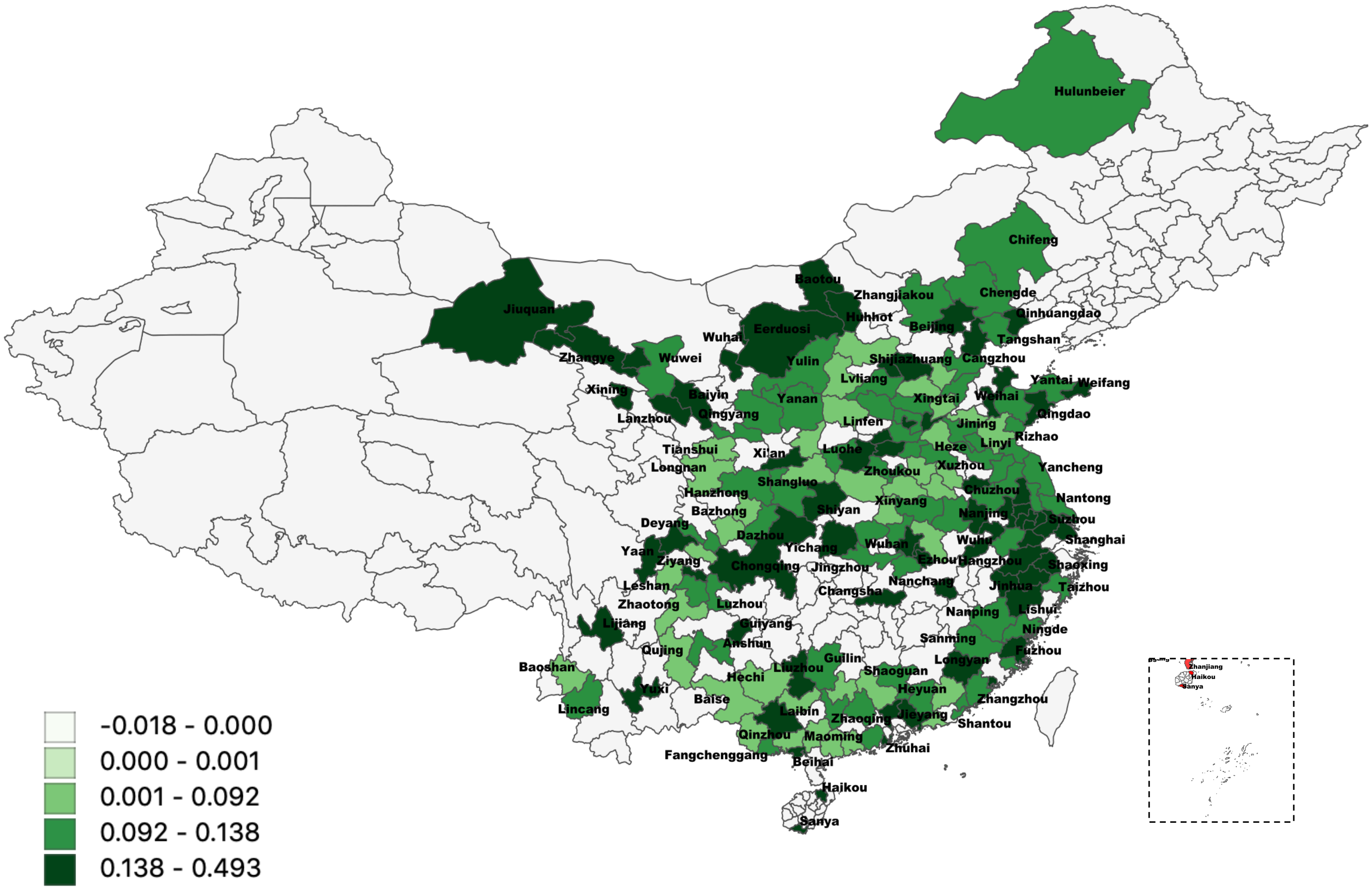
Top 5

- Zhengzhou 郑州
- Shenzhen 深圳
- Guangzhou 广州
- Chengdu 成都
- Changsha 长沙

Bottom 5

- Ziyang 资阳
- Weinan 渭南
- Shangluo 商洛
- Wuwei 武威
- Hulunbeir 呼伦贝尔

Prediction Results | 10 years growth in China



Top 5

- Zhengzhou 郑州
- Sanya 三亚
- Haikou 海口
- Zhuhai 珠海
- Guangzhou 广州

Bottom 5

- Jieyang 揭阳
- Lvliang 吕梁
- Weinan 渭南
- Zhaotong 昭通
- Zhoukou 周口

1st Report – Summary

Predicting City Growth with Machine Learning

- **Information transmission, computer service and software & Finance and Insurance were the two fastest-growing industries in China during the past 15 years**

- **Healthcare & Professional, scientific and technical services were the two fastest-growing industries in US during the past 15 years**

- **Zhengzhou, Guangzhou, Shenzhen, Zhuhai are predicted to be the fastest growing cities in the next 5-10 years**

Machine learning offers a set of methods that outperform OLS in terms of out-of-sample prediction.

But: in most cases, ML methods are not directly applicable for research questions in econometrics and allied fields, especially when it comes to causal inference.

How can we exploit the strengths of supervised ML (automatic model selection & prediction) for causal inference?

City Growth Projection | Regression Model

Changes on Changes urban growth regression

$$\Delta_{t+1,t} \log N_i = \beta_0 - \beta_I \Delta_{t+1,t} \log D_i + \epsilon_{it}$$

Supposing the myopic adjustment process $\mathrm{N}_{it+1} = N_i^{*\lambda} N_{it}^{1-\lambda}$, where N_i^* denotes the equilibrium steady-state population, we can interpret λ as the rate of convergence.

If $\lambda = 0$, there is no mobility, and if $\lambda = 1$, the population adjustment is immediate.

Taking logs and readjusting yields:

$$\Delta_{t+1,t} \log N_i = \lambda(\log N_i^* - \log N_{it})$$

According to the spatial equilibrium condition, DN_i^* must be constant in steady state. Thus

$$\log N_i^* = \beta_0 - \beta_1 \log D_{it}$$

Estimation methods

- OLS: include all regressors and minimizes mean square errors.
- LASSO (Tibshirani 1996) with penalty level (lambda) selected by information criteria EBIC (Chen and Chen 2008): The lasso minimizes the residual sum of squares (RSS) subject to a constraint on the absolute size of coefficient estimates.
- Square-root LASSO (Belloni, Chernozhukov, and Wang 2011, 2014): The sqrt-lasso is a modification of the lasso that minimizes (RSS)^(1/2) instead of RSS.
- Both of these LASSO methods use CV.