Predicting neighborhoods' socioeconomic attributes using restaurant data [1]

Lei Dong

MIT
Senseable City Lab & China Future City Lab

leidong@mit.edu
donglei.org

August, 2019

[1] These slides are based on a paper published at *PNAS* by Lei Dong, Carlo Ratti, and Siqi Zheng. doi.org/10.1073/pnas.1903064116

Abstract

- High resolution, timely socioeconomic data (e.g., population, employment, enterprise activity) are foundation for many fields, including urban planning, economic monitoring, and enterprise's location choice.
- In many developing countries/cities, **reliable local scale socioeconomic data remain scarce**.
- Here we show an easily accessible and timely updated location attribute – **restaurant** – can be used to accurately predict a range of socioeconomic attributes.
- We combined the restaurant data from an online platform with three novel micro-datasets in nine cities. We train machine learning models to estimate **daytime/nighttime population, number of firms, and consumption level** at various spatial resolution.
- We demonstrate the cross-city generality of this method by training model in one city and then applying it to other cities, as well as showing the importance (explanation) of variables.

Satellite imagery and poverty prediction

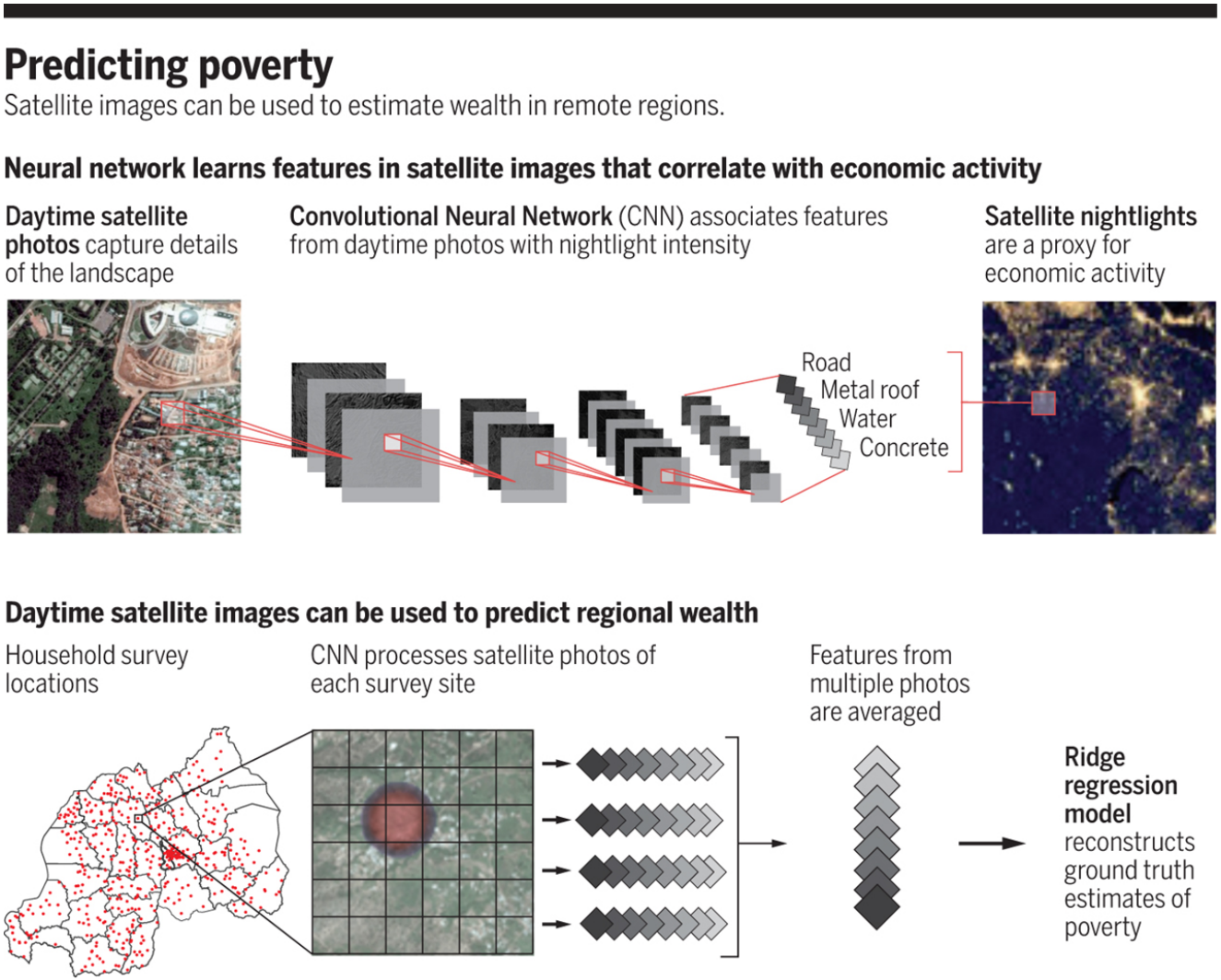


Figure 1: Blumenstock, 2016; Jean et al., 2016

Street view imagery and urban change prediction

Figure 2: Naik et al., 2017

Literature

- Remote sensing data (e.g. nighttime light), street view imagery
  - Nighttime light and econoimc growth (Henderson, Storeygrad and Weil, 2012, *AER*)
  - Satellite imagery poverty prediction (Jean et al., 2016, *Science*)
  - Satellite data in econmmics (Donalson and Storeygard, 2016, *JEP*)
  - Street view and physical urban change (Naik et al., 2017, *PNAS*), demographic makeup of neighborhoods (Gebru et al., 2017, *PNAS*)
- Mobile phone data, social media data
  - Mobile phone data poverty prediction (Blumenstock et al., 2015, *Science*)

Literature

- Restaurant data (Yelp)
  - Using Yelp data to measure economic activity (Glaeser et al., 2017, *NBER Working Paper*), neighborhood change (Glaeser et al., 2018, *AEA Papers and Proceedings*); Hygiene inspections (Kang et al., *EMNLP, 2013*)
- Machine learning and econometric
  - Mullainathan and Spiess, 2017, *JEP*; Einav and Levin, 2014, *Science*; Varian, 2014, *JEP*

Facebook *Like* Predicts Personality

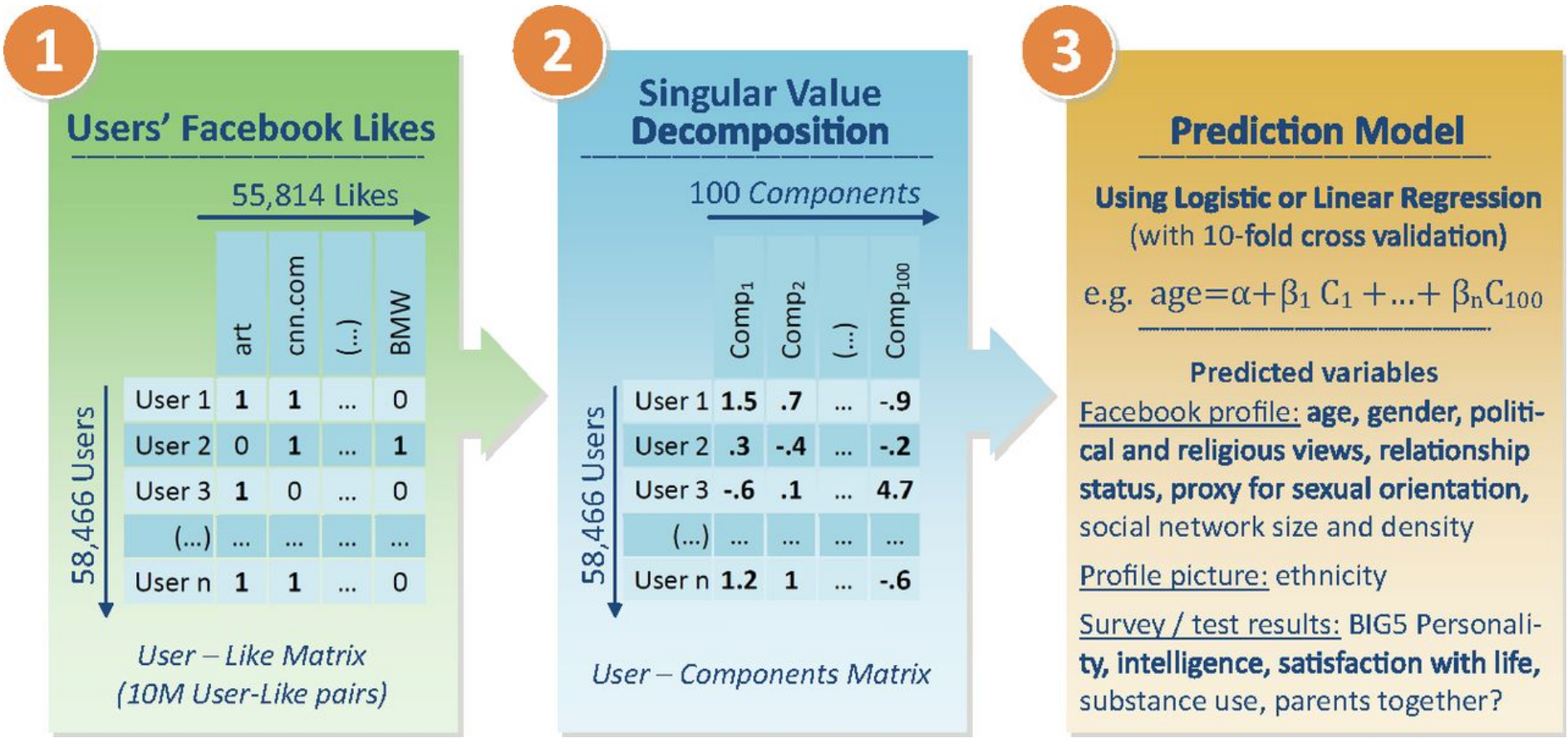


Figure 3: Michal Kosinski et al., 2013

Facebook *Like* Predicts Personality

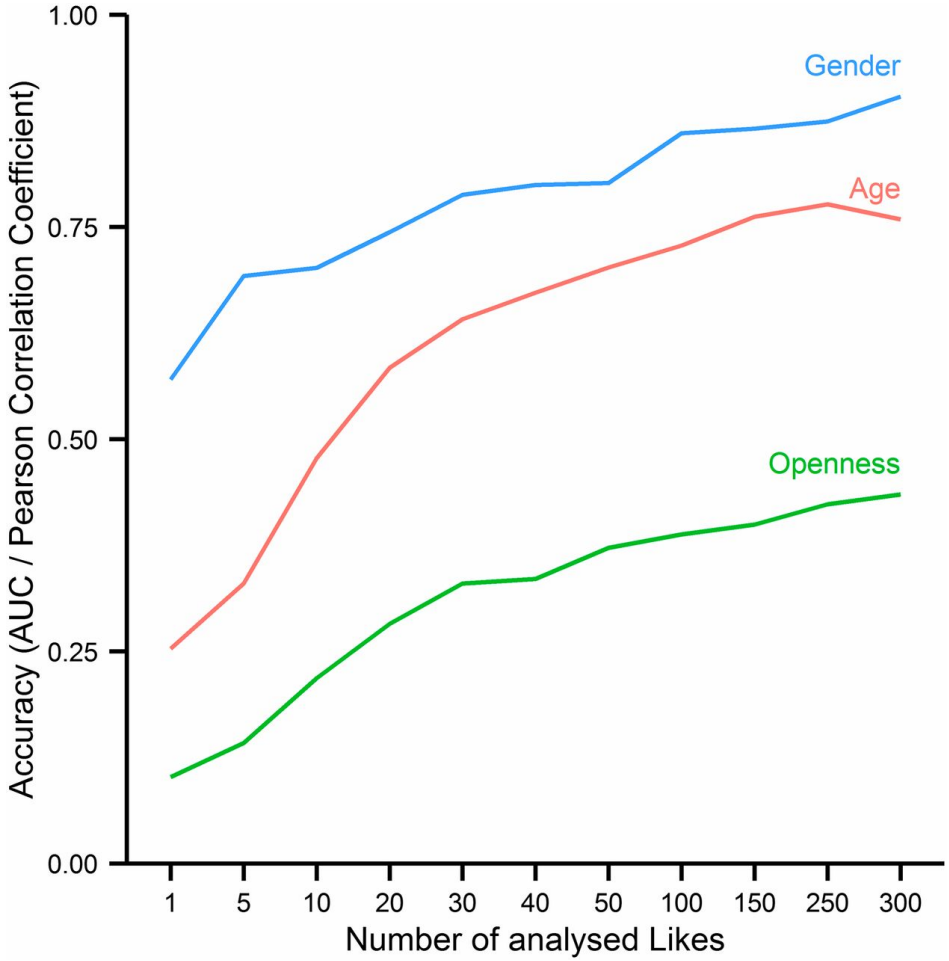


Figure 4: Michal Kosinski et al., 2013

Research Question

How it works out that restaurant data on estimating socioeconomic variables?

Advantages of Dianping Restaurant Data

Compared with mobile phone data, remote sensing, street-view imagery, social media, Restaurant data is:

- Easy access
- Update in a timely manner
- Well-structured
- Nation-wide coverage
- Clear economic meaning

Could Dianping *Restaurant* Predicts Urban Vibrancy?

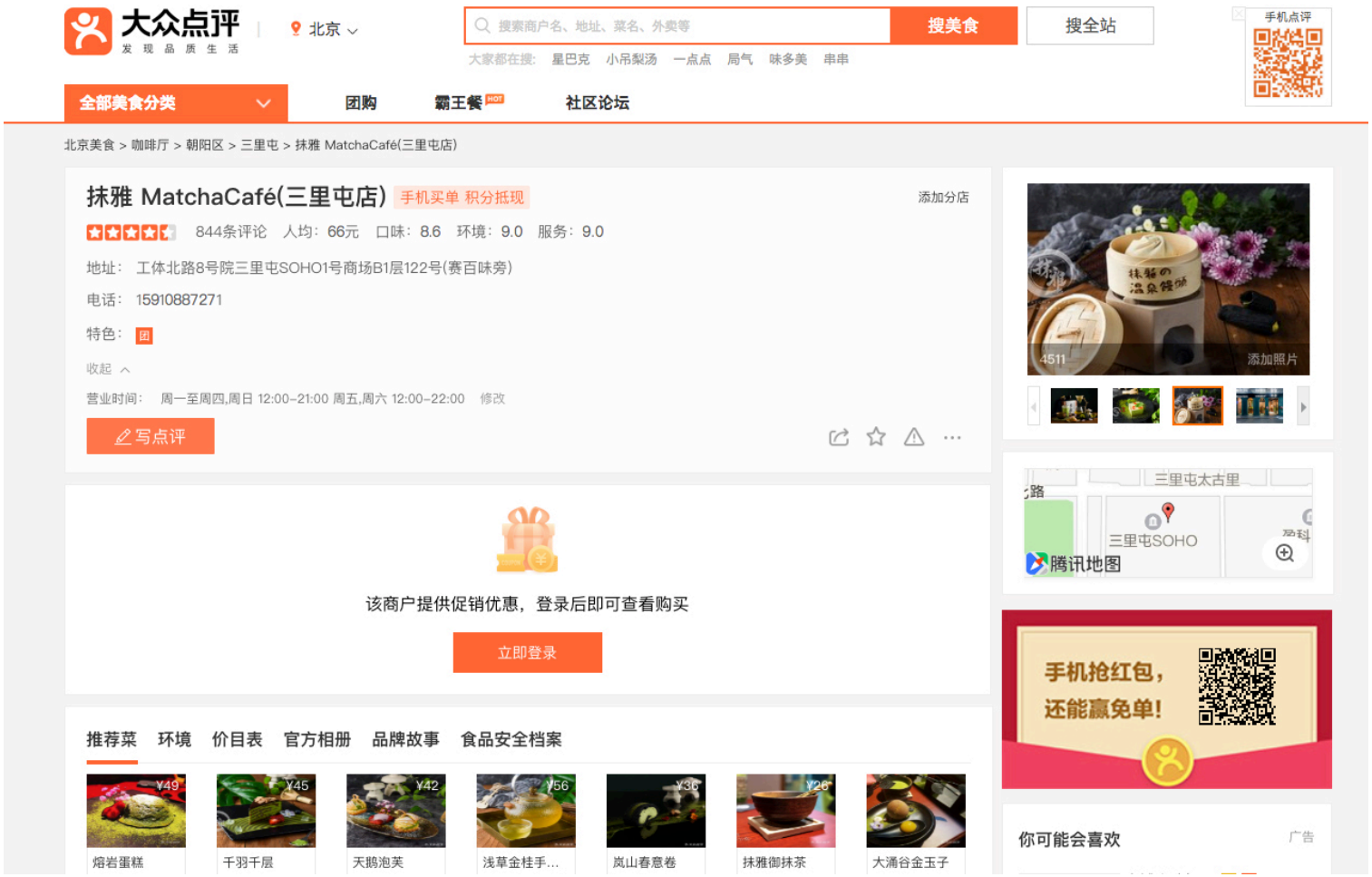


Figure 5: Shop of Dianping.com

Could Dianping *Restaurant* Predicts Urban Vibrancy?

Figure 6: Spatial Distribution of Restaurants in Beijing

Research Framework

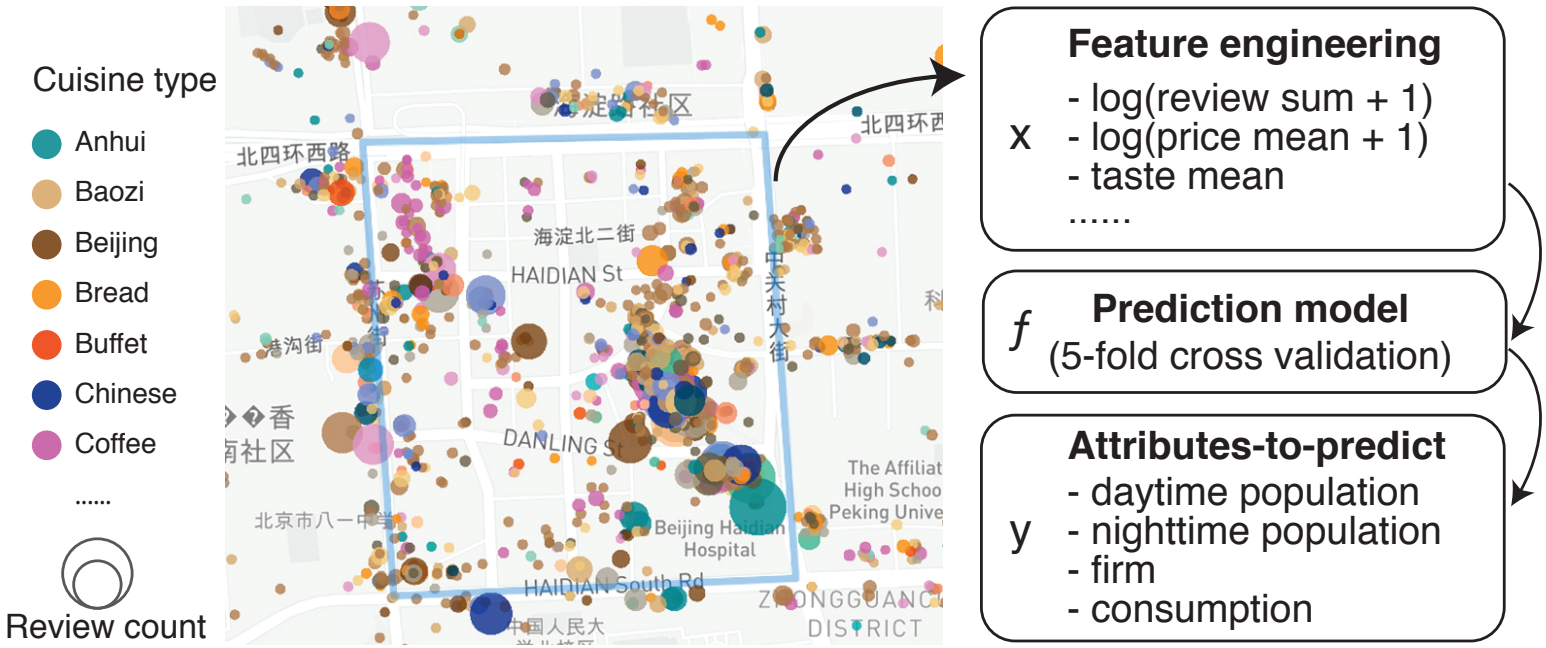


Figure 7: A schematic of feature engineering and the training process. For each grid cell and for each cuisine type of restaurant, we calculate seven metrics. We merge restaurant features with attributes-to-predict (daytime population, nighttime population, firm, and consumption) by grid cell index.

Research Framework

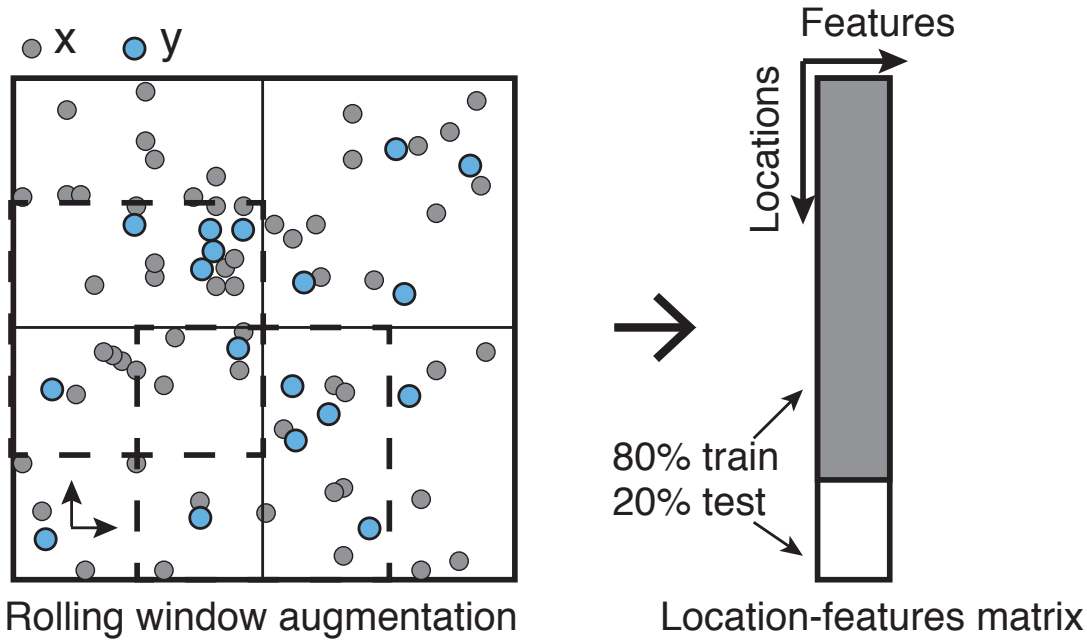


Figure 8: Rolling window data augmentation.

- Rolling grid window to get more data, and this is also helpful to reduce possible bias caused by the arbitrary partition of grid cells.
- Very helpful to improve the accuracy of machine learning models

Attributes of Restaurants

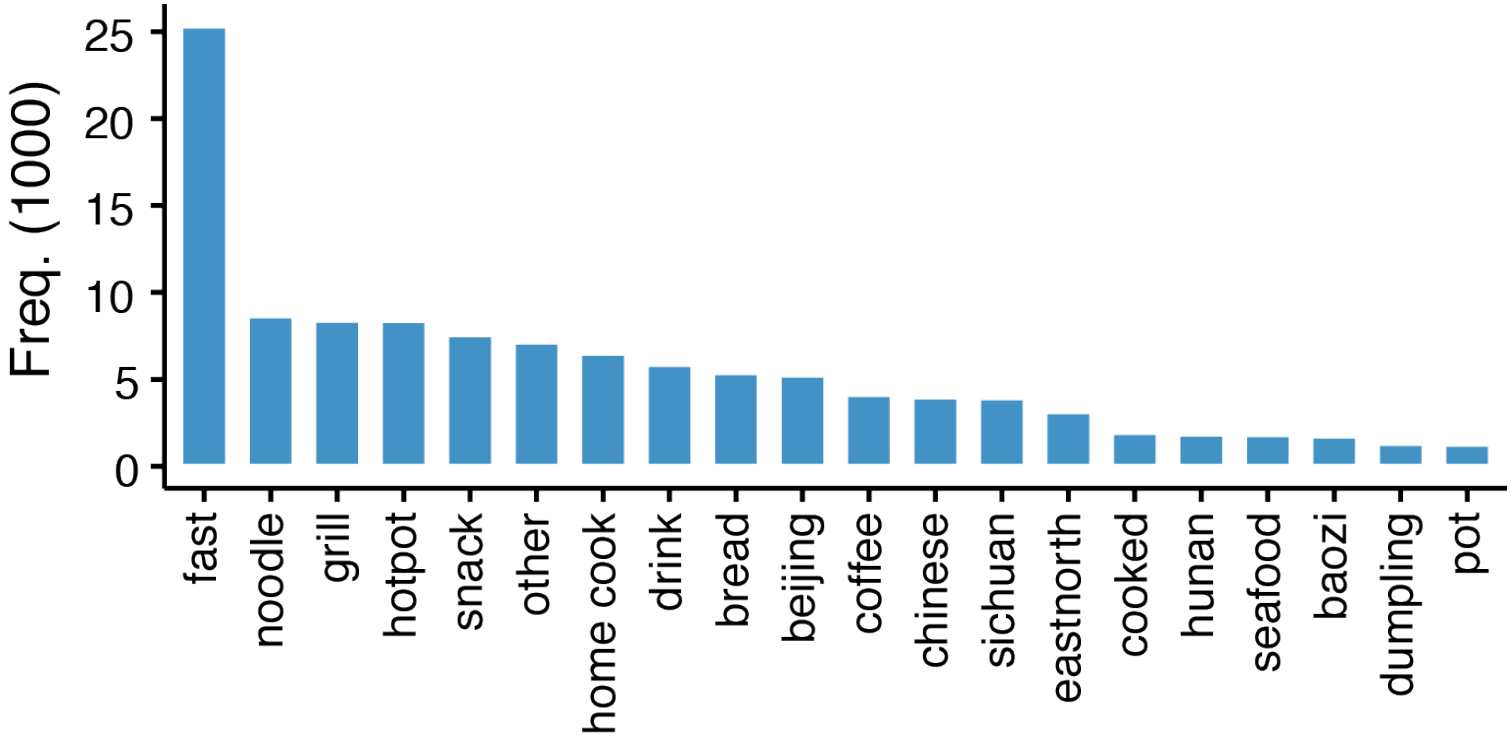


Figure 9: Distribution of cuisines (Beijing)

Socioeconomic Attributes to Predict

All the following data are not easy to access in a fine spatial resolution and timely manner.
First stage:

- Population (estimated by mobile phone data)
- Employment (estimated by mobile phone data)
- Number of firms (Industrial and Commercial Bureau)
- Consumption (estimated by POS (point-of-sale) data)

Second stage:

- **Heterogeneity** of predicting accuracy within the city
- **Transferability**: Could we use the model trained in big cities to infer attributes in small cities?

Cities

Name	Province	Urban Popu. (million)	GDP (billion $)	Built-up Area (km^2)	# Restaurants (1,000)
Beijing	-	21.71	414.71	1419.66	143.96
Shenzhen	Guangdong	12.53	332.33	923.25	104.00
Chengdu	Sichuan	11.53	205.70	837.27	138.63
Shenyang	Liaoning	6.73	115.10	588.26	58.13
Zhengzhou	Henan	4.70	117.39	422.35	70.62
Kunming	Yunnan	3.97	63.71	435.81	58.78
Baoding	Hebei	2.84	52.99	185.70	46.25
Yueyang	Hunan	1.27	46.34	97.00	17.57
Hengyang	Hunan	1.18	41.77	113.53	19.93

Table 1: Basic statistics of each city (Data source: wikipedia.org, dianping.com, China City Statistical Yearbook)

Machine Learning Model

- LASSO (least absolute shrinkage and selection operator)

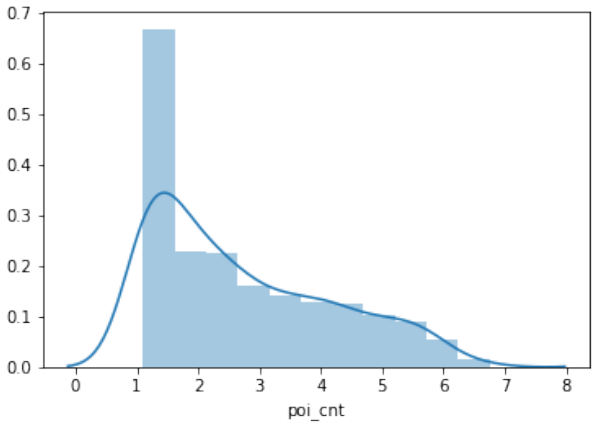
$$\min_{\beta_0,\beta_1} \frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda\left[(1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right]$$

  where λ is the penalty parameter, $\alpha = 1$ (LASSO) and $\alpha = 0$ (Ridge)
- 80% training and 20% for testing
- 5-fold validation
- In the setting of our search question, LASSO is more generalizable (compared with other machine learning models) because it can shrink coefficients.

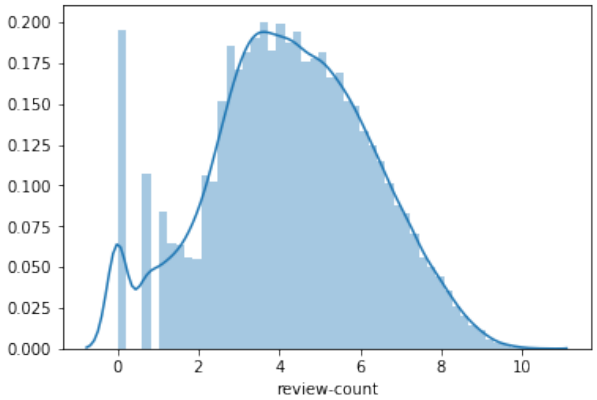
Features and Data Sampling

- Aggregate at the grid cell level (1km, 2km, ...)
- For each grid cell and for each cuisine type of restaurant, calculating:
  - log(restaurant count + 1)
  - log(review sum + 1)
  - log(review mean + 1)
  - log(review median + 1)
  - log(review no-zero + 1)
  - log(average price + 1)
  - taste mean
  - environment mean
  - service mean
  - star mean
- Add the average of eight neighbourhood grid cells as new features
- Merge with *attributes-to-predict* by grid cell id

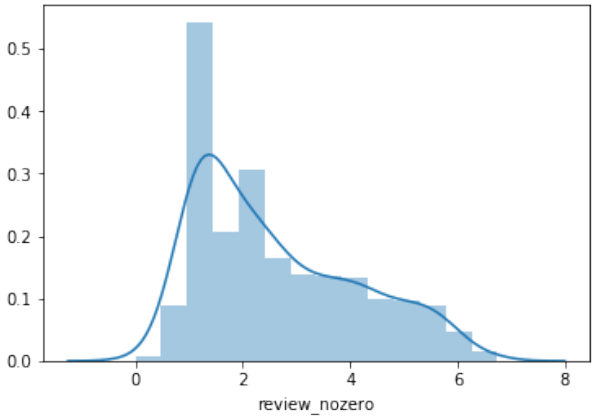
Features and Data Sampling



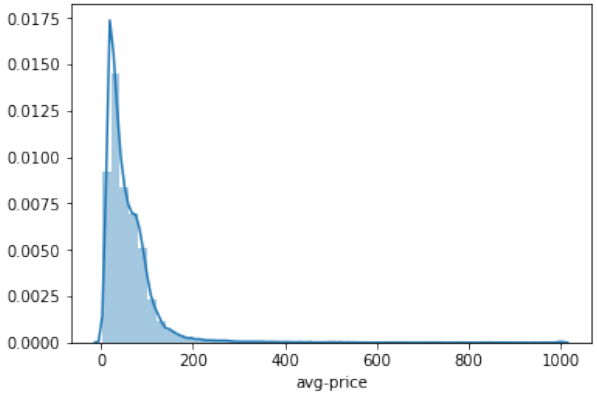
(a) log(poi count + 1)



(b) log(review count + 1)



(c) log(review nozero + 1)



(d) Average price

Prediction Performance

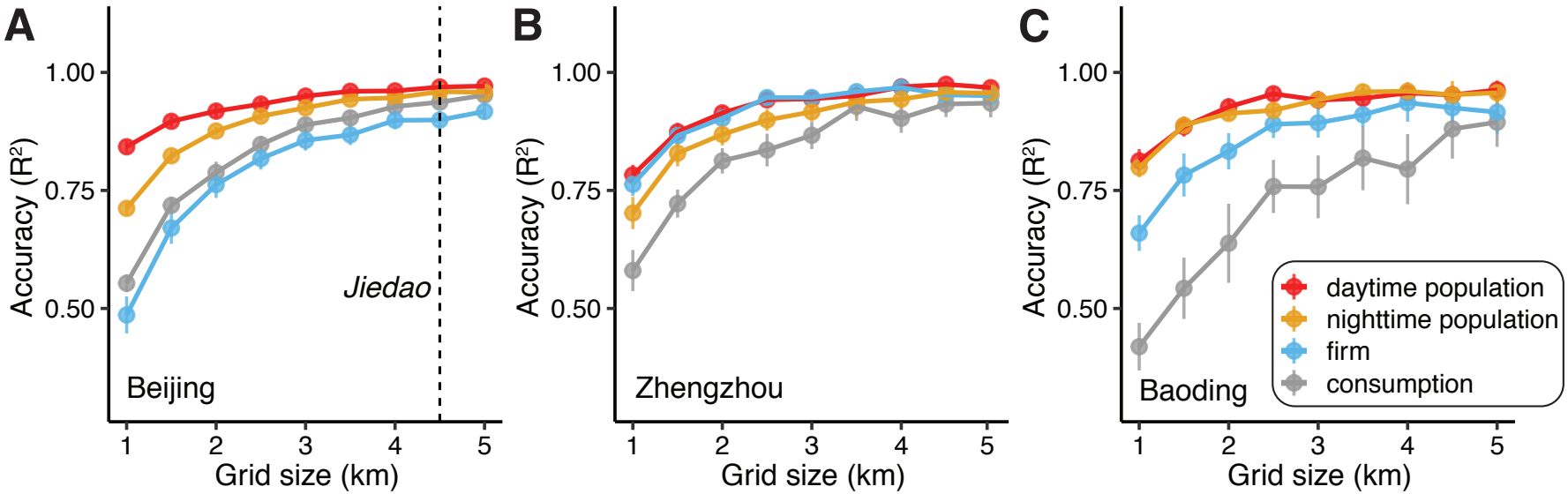


Figure 11: **Prediction accuracy.** (A) Beijing. (B) Zhengzhou. (C) Baoding. The red, yellow, gray and blue lines represent the accuracy of daytime population, nighttime population, firms, and consumption, respectively.

Prediction Performance

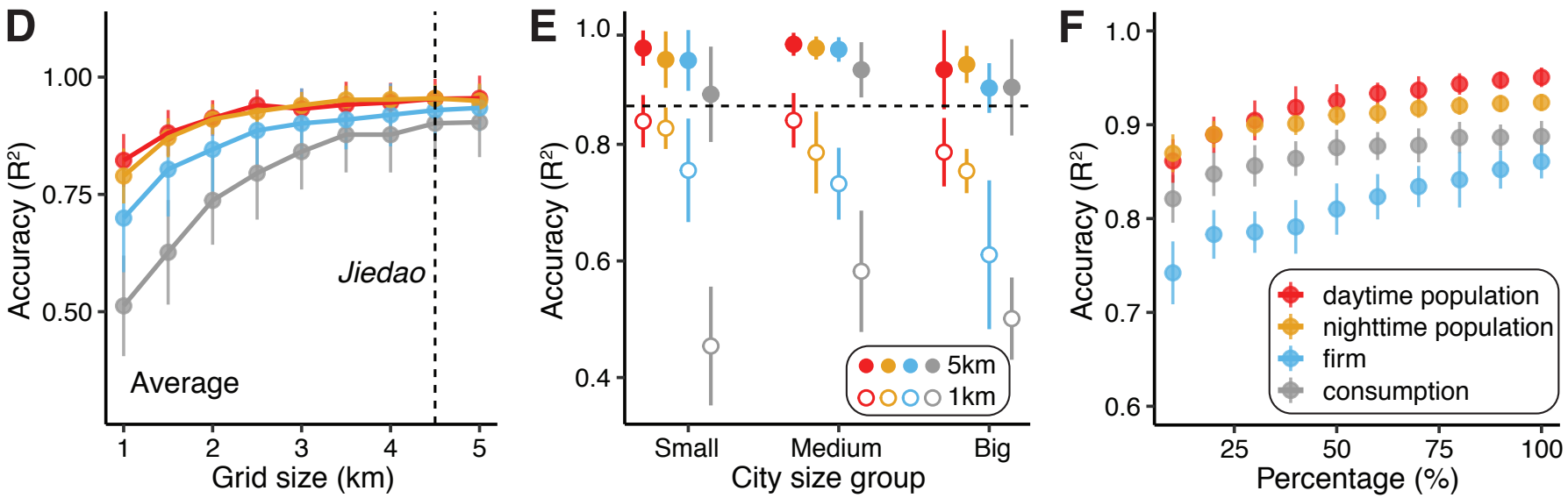


Figure 12: **Prediction accuracy.** (D) Averaged accuracy of nine cities. (E) The relationship between city size and model accuracy. Big cities are cities with a population of over 10 million; Medium-sized cities have a population of 3-10 million; Small-sized cities have a population of less than 3 million. (F) The percentage of training samples and accuracy. Models were trained from data of Beijing at 3km resolution.

Prediction Performance

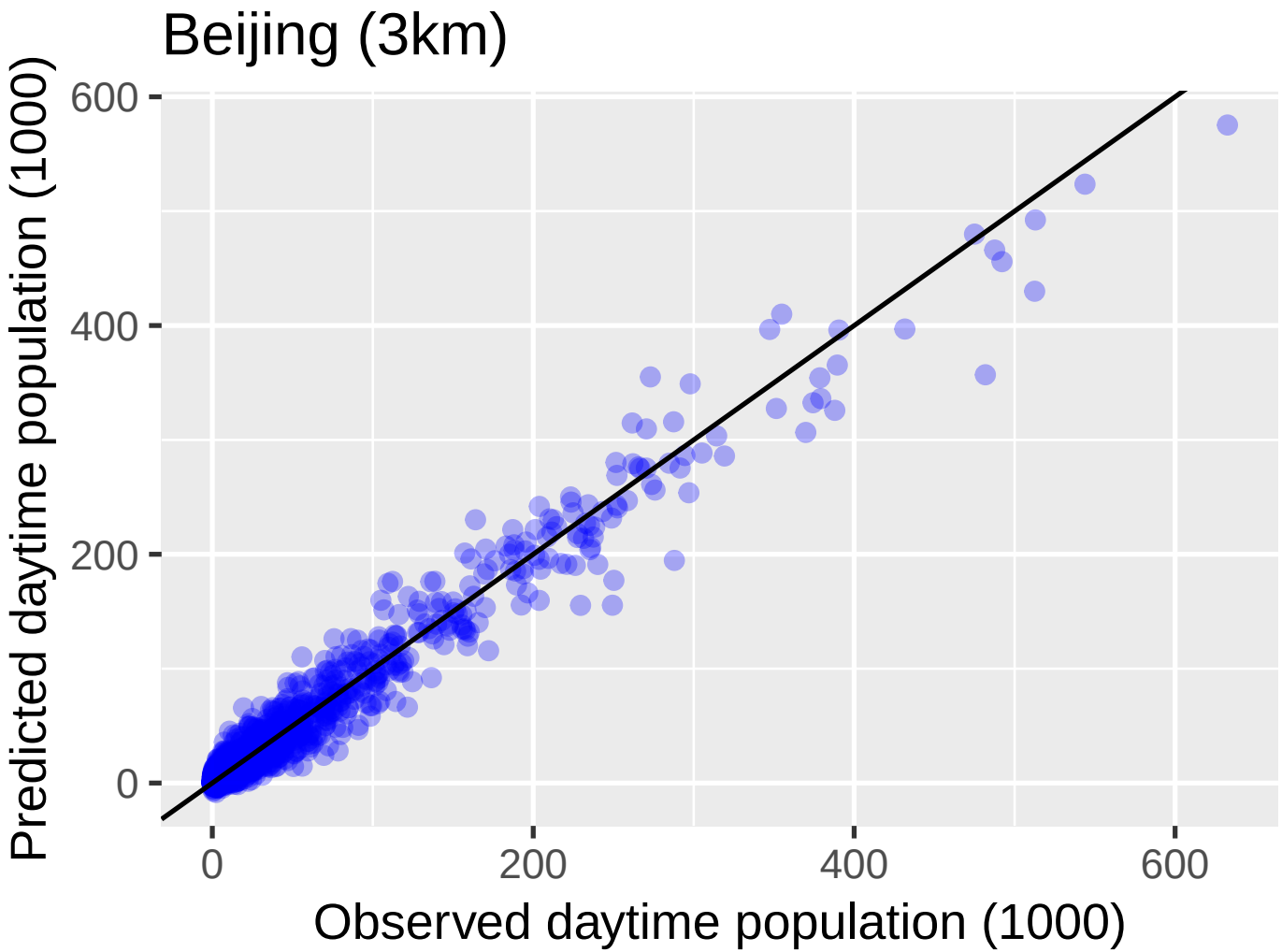


Figure 13: Prediction accuracy. (Daytime population)

Heterogeneity

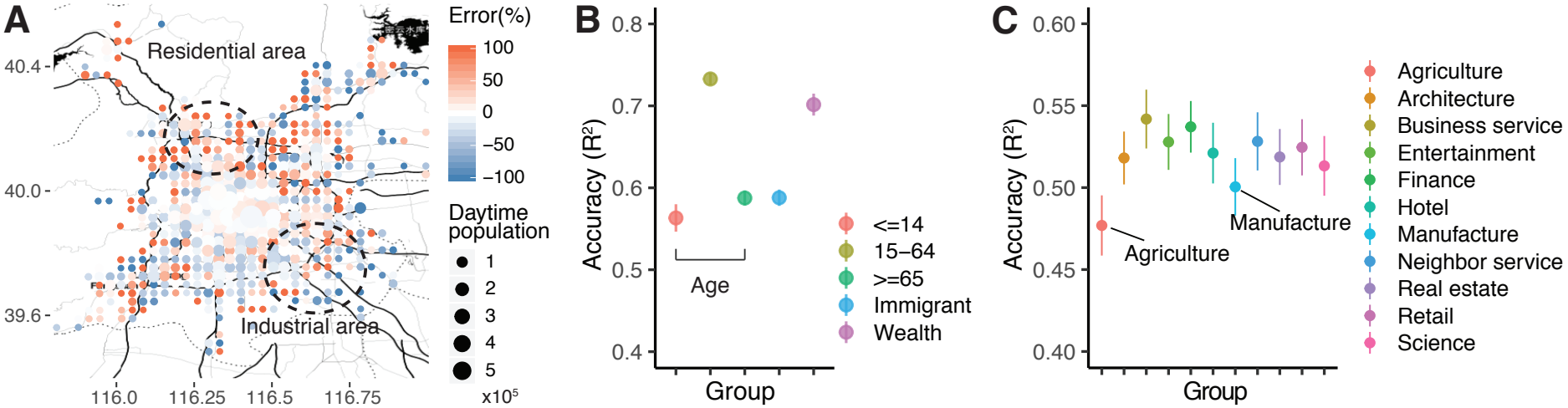


Figure 14: Heterogeneity. (A) Spatial distribution of prediction errors of daytime population at 3km resolution (Beijing). (B) Prediction accuracy of different age groups, immigrant percentage, and wealth (housing price) at the Jiedao level. (C) Prediction accuracy of different industries of firms at the Jiedao level.

Top Predictors

- The best predictors of the presence of immigrants include cuisines of "Beijing", "Bakery", "Cooked", and "Xinjiang";
- Whereas housing price (wealth) is indicated by "Hubei", "Coffee", and "Beijing".
- Good predictors of the 15-64 age group include "Bakery", "Hotpot", and "Seafood".
- Although some restaurant features clearly relate to their predicted attribute, as in the case of "Coffee" for the housing price prediction, other pairs are more elusive; there is no obvious connection between "Bakery" and presence of immigrants.

Transfer

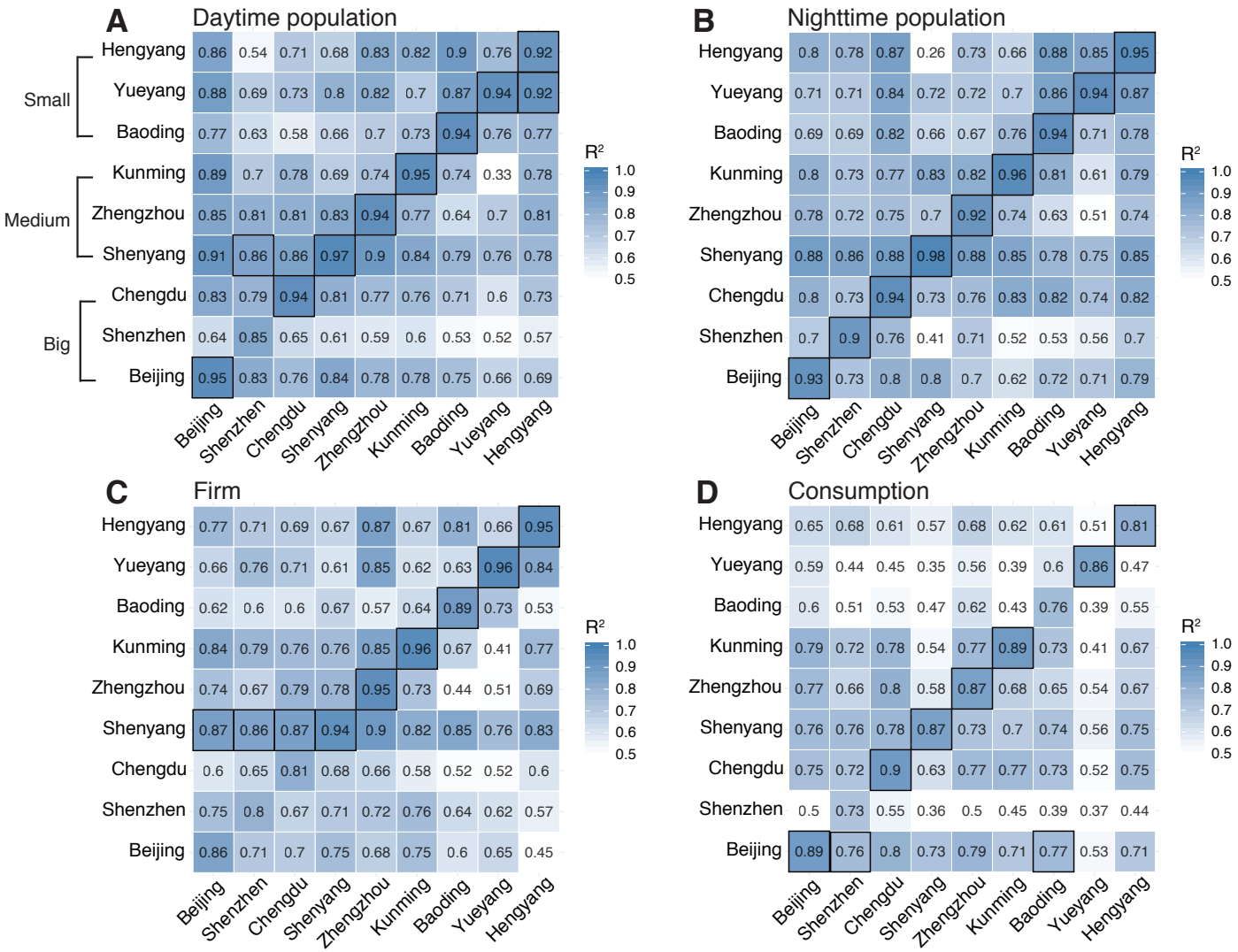


Figure 15: Cross-city model generalization. Cross-validated R^2 of models trained in one city (x axis) and applied to other cities (y axis).

With and Without Data Augmentation

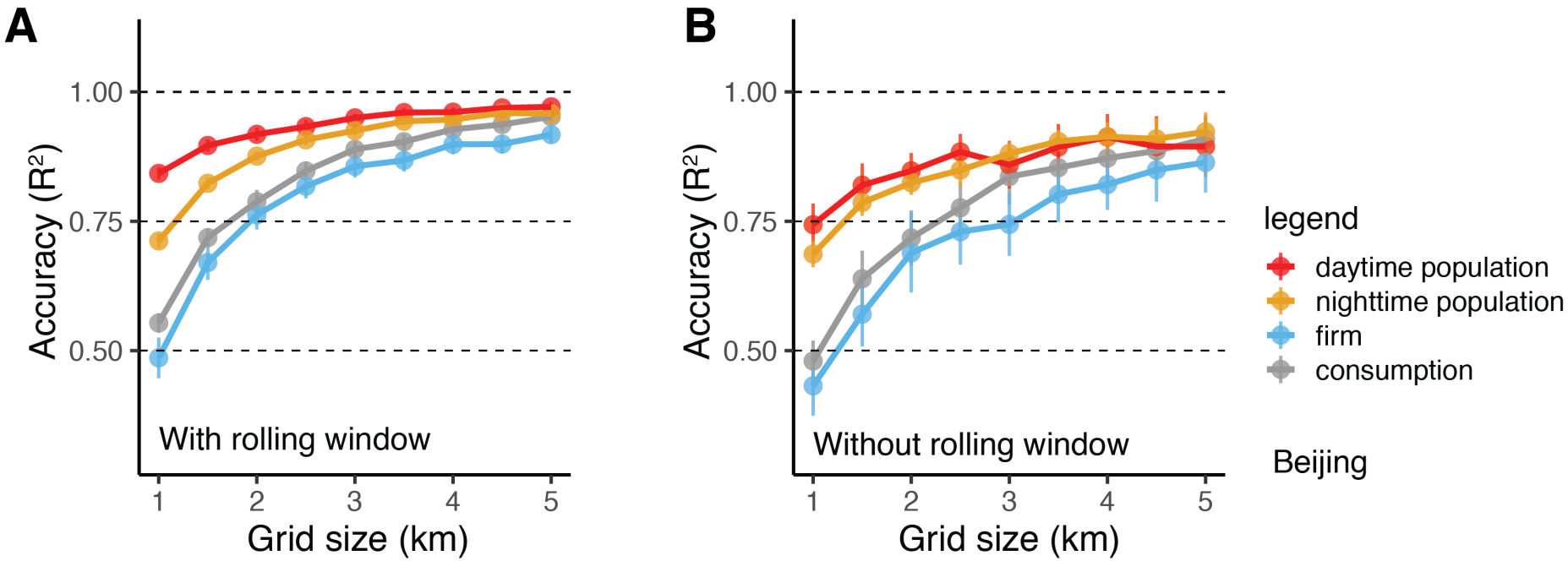


Figure 16: Comparison between with and without rolling window data augmentation in Beijing. (A) With rolling window. (B) Without rolling window. Rolling window augmentation significantly improves model performance.

Restaurant vs. Nighttime Light

Cell size (km)	Daytime Popu.		Nighttime Popu.		# Firm		Consumption	
	Rest.	Light	Rest.	Light	Rest.	Light	Rest.	Light
1.0	0.842	0.741	0.712	0.480	0.486	0.287	0.553	0.382
1.5	0.896	0.781	0.823	0.591	0.670	0.380	0.718	0.517
2.0	0.918	0.821	0.876	0.667	0.762	0.454	0.788	0.597
2.5	0.933	0.826	0.907	0.772	0.818	0.540	0.847	0.638
3.0	0.950	0.844	0.925	0.810	0.856	0.611	0.889	0.685
3.5	0.960	0.853	0.943	0.834	0.868	0.629	0.904	0.704
4.0	0.961	0.866	0.946	0.843	0.899	0.642	0.928	0.768
4.5	0.969	0.868	0.959	0.862	0.899	0.684	0.937	0.764
5.0	0.971	0.868	0.958	0.886	0.917	0.688	0.952	0.749

Table 2: Prediction performance (R^2) of restaurant model (Rest.) and nighttime light model (Light) of Beijing

Conclusions

- We demonstrate that local restaurants can accurately infer the spatial distribution of socioeconomic activities within cities at high granularity.
- We also show that collecting only a few training samples can result in high accuracy in inferring unsampled locations using machine learning models, suggesting that the restaurant model can help city governors and researchers monitoring city performance in a timely and low-cost manner.
- Despite differences in geographical, cultural and economic conditions, cities share many common features in restaurants, which are strongly correlated with socioeconomic characteristics across cities. The transferability of the model could help bridge the "socioeconomic data gap" between large and small cities.

Conclusions

- Given the limited availability of high-resolution time series data for key socioeconomic indicators, we have not yet been able to evaluate the ability of the restaurant data and the machine learning approach to predict the temporal changes in a location's socioeconomic attributes over time.
- Another untouched but very important direction is deriving high granular data from coarse aggregated sources like census data or other survey instruments. (e.g., super-resolution in computer vision)
- Big data and machine learning are playing important roles in the urban agenda. But "there are a number of gaps between making a prediction and making a decision." (Athey, 2017)
- This is especially important for cities, since most decisions about city development are long term decisions, while the current predictive models are trained using short term data. How to use these data to assist decision-making requires more in-depth research and practice.

Thank You