Predicting neighborhoods' socioeconomic attributes using restaurant data

Lei Dong[a,b], Carlo Ratti[a], and Siqi Zheng[b,1]

[a]Senseable City Lab, Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]China Future City Lab and Center for Real Estate, Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by William A. V. Clark, University of California, Los Angeles, CA, and approved June 10, 2019 (received for review February 21, 2019)

Accessing high-resolution, timely socioeconomic data such as data on population, employment, and enterprise activity at the neighborhood level is critical for social scientists and policy makers to design and implement location-based policies. However, in many developing countries or 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 of urban neighborhoods. We merge restaurant data from an online platform with 3 microdatasets for 9 Chinese cities. Using features extracted from restaurants, we train machine-learning models to estimate daytime and nighttime population, number of firms, and consumption level at various spatial resolutions. The trained model can explain 90 to 95% of the variation of those attributes across neighborhoods in the test dataset. We analyze the trade-off between accuracy, spatial resolution, and number of training samples, as well as the heterogeneity of the predicted results across different spatial locations, demographics, and firm industries. Finally, we demonstrate the cross-city generality of this method by training the model in one city and then applying it directly to other cities. The transferability of this restaurant model can help bridge data gaps between cities, allowing all cities to enjoy big data and algorithm dividends.

socioeconomic data | restaurant | urban studies | machine learning | social good

High-resolution socioeconomic data for cities are critical for researchers and policy makers. Such data, like spatial–temporal distribution of population and economic activity, provide essential guidance for researchers in urban, economic, and environmental fields as they seek to understand city activities (1). The main source of socioeconomic data are various types of surveys. For instance, population census provides comprehensive demographic information, while household consumption survey allows us to understand changes in the consumption structure. However, because of the high cost, surveys cannot consider both high spatial resolution and timeliness—a census is typically conducted once every 10 y. More importantly, in some developing countries such as China, key socioeconomic data are often scarce or of low quality. Even in some big Chinese cities like Beijing and Shanghai, most socioeconomic databases are only publicly accessible at the district level—offering fewer than 20 spatial observations.

Given the lack of timely and spatially detailed socioeconomic data, researchers and policy makers seek alternative approaches to measuring socioeconomic outcomes of interest (2, 3). "Nighttime lights" data, for instance, has contributed to estimating regional economic activities (4–6). Massive online data, generated by social media, mobile phone, and e-commerce platforms, have been used to infer individual's personality (7), unemployment (8–10), population distribution (11), wealth (12), consumption index (13), and so on. Moreover, current progress with machine-learning algorithms has also made "unstructured data" (e.g., satellite/street-view imagery and text content) valuable in inferring socioeconomic outcomes (14–16). Nevertheless, accessibility, updatability, and interpretability of these data sources remain significant challenges.

In this paper, we show the potential of using restaurant data to predict 4 important socioeconomic variables: daytime population, nighttime population, number of firms, and volume of consumption at various spatial resolutions. The restaurant data we use are a set of attributes, such as average price of a meal, cuisine category, and so on, that characterize a restaurant. This exercise has 3 motivations. First, China has experienced rapid urbanization over the past decades, but fine granular and regularly updated urban socioeconomic data are rarely available. Second, the restaurant industry is one of the most decentralized and deregulated local (location-based) consumption industries, whether in developed or developing cities. It is highly correlated with local socioeconomic attributes, like population size, wealth, and consumption. Third, national-wide restaurant data are easily available via online platforms such as Dianping (China) and Yelp (United States), and variables extracted from restaurant data (e.g., average price, number of reviews) have reasonable economic interpretation, providing the possibility of explaining model's results.

This paper differs from recent literature that has evaluated the relation between restaurant data from Yelp and US county business patterns (17), neighborhood change (18), segregation (19), or hygiene quality (20). We combine restaurant data with 3 extensive datasets—population, firm, and consumption—and we test prediction accuracy at different spatial resolutions, the heterogeneous effect, as well as the cross-city generality by training the model in one city and then applying it to other cities. Our

Significance

High-resolution socioeconomic data are crucial for place-based policy design and implementation, but it remains scarce for many developing cities and countries. We show that an easily accessible and timely updated neighborhood attribute, restaurant, when combined with machine-learning models, can be used to effectively predict a range of socioeconomic attributes. This approach allows us to collect training samples from representative neighborhoods and then use our trained model to infer unsampled neighborhoods in the city in a granular, timely, and low-cost manner. The good cross-city transferability performance of our model can also help bridge the "data gap" between cities, by training the model in cities with rich survey data and then applying it to cities where such data are unavailable.

Author contributions: L.D., C.R., and S.Z. designed research; L.D. performed research; L.D. analyzed data; and L.D., C.R., and S.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The aggregated data and code, to replicate results in this paper, have been deposited in GitHub, https://github.com/leiii/restaurant.

[1]To whom correspondence may be addressed. Email: sqzheng@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1903064116/-/DCSupplemental.

Published online July 15, 2019.

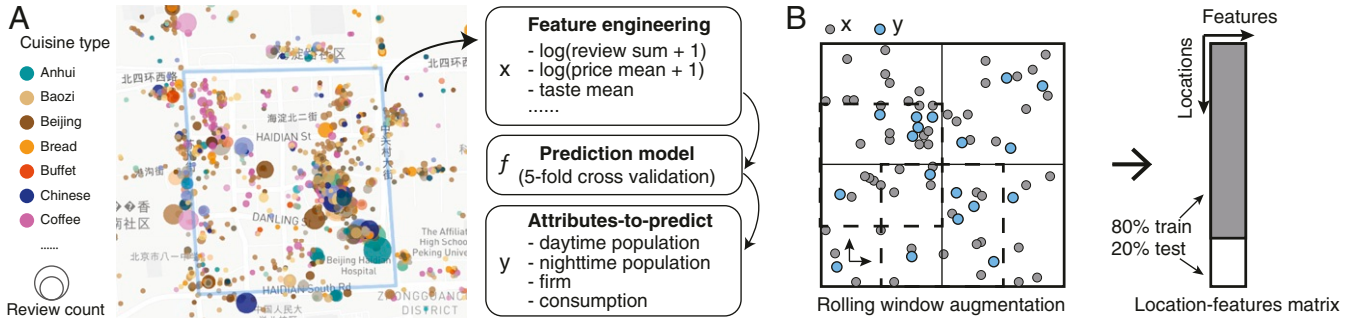

Fig. 1. Data and methodology. (*A*) A schematic of feature engineering and the training process. For each grid cell and for each cuisine type of restaurant, we calculate 7 metrics (see *SI Appendix*, Note 2 for details). We merge restaurant features with attributes-to-predict (daytime population, nighttime population, firm, and consumption) by grid cell index. (*B*) Rolling window data augmentation. For each grid, we move the sampling window along the *x* and *y* direction (and both) by half the length of the grid size to produce a new sample that is 3 times the original sample size, and then we split the combined data into the 80% training set and the 20% test set.

approach is especially helpful for developing countries/cities with a fast speed of urbanization and limited resources: 1) it allows us to collect training samples of socioeconomic variables from a small number neighborhoods and then infer unsampled neighborhoods from statistical models. 2) The transferability of the model can help bridge the "data gap" between big and small cities, allowing small cities to benefit from data and algorithm advances.

Results

For this research, we collect restaurant data from Dianping (the largest online rating and deal service platform for restaurants in China) for 9 Chinese cities: Beijing, Shenzhen, Chengdu, Shenyang, Zhengzhou, Kunming, Baoding, Yueyang, and Hengyang. *SI Appendix*, Table S1 shows the descriptive statistics of these cities. These selected cities are located in different geographic regions of China and vary greatly in population size, which helps test the robustness of our method. We collect the following restaurant attributes: name, longitude, latitude, cuisine category, price, taste rating, environment rating, service rating, average score, and number of reviews (*SI Appendix*, Fig. S1). From these restaurant attributes, we construct hundreds of features, for instance, the number of restaurants, the number of reviews, the mean of reviews, and the mean of taste ratings at various spatial resolutions—from 1 to 5 km (see Fig. 1*A* and *SI Appendix*, Note 2 for details). We then merge restaurant features with daytime/nighttime population data (estimated by mobile phone data), firm registration record data, and volume of consumption (estimated by the bank card records) for each grid cell (*Materials and Methods*). We propose a data augmentation method to improve model performance. For each grid cell, we move the sampling window along the *x* and *y* direction by half the length of the grid size (Fig. 1*B*). This data augmentation method provides more information about the spatial distribution of variables and could significantly improve the accuracy and stability of the model (*SI Appendix*, Fig. S7). For each city and each spatial resolution, we apply LASSO (least absolute shrinkage and selection operator) regressions (21) with 5-fold cross-validations on 80% of the observations (training set) and then test the model performance on the remaining 20% of the sample (testing set) (see *Materials and Methods* for details).

Model Performance. Fig. 2 and *SI Appendix*, Fig. S8 show the prediction accuracy (R^2) of daytime population, nighttime population, number of firms, and volume of consumption across 9 cities at various spatial resolutions (from 1 to 5 km). In general, the model trained by restaurant data are strongly predictive of all 4 variables of interest. At a cell size of 4.5 km (the resolution of the administrative boundary—Jiedao—in China, similar to the neighborhood level in the United States) predictions on the test dataset can explain 95% (minimum: 86%; maximum: 98%) of the variation in daytime population, 95% (91%; 98%) in nighttime population, 93% (87%; 97%) in number of firms, and 90% (82%; 94%) in the volume of consumption (Fig. 2*D*). The highest accuracy is achieved for daytime/nighttime population, which are proxies for employment and residents, respectively (22), suggesting that as a local market-driven industry, restaurants are highly correlated with characteristics of the local population. The result of population estimation (95%) is also much higher than remote sensing-based models at a similar granular level (R^2 of the predicted population density reported in ref. 23 was 86% and 85% in ref. 11).

The R^2 values of firm prediction are a little bit lower than for daytime/nighttime population. This may be attributable to the limitation of the firm dataset. For firm data, we only have the registered address, which may not be the same as the operation address, and firm size is unreported in the raw dataset and thus may also bias the results. For the consumption side, restaurants should highly correlate with food consumption, while the category of consumption is unknown in the dataset, making it difficult to test this hypothesis.

Fig. 2 *A*–*D* also shows that accuracy increases with the size of the grid. This may be because an increase in cell size reduces heterogeneity among cells, thus reducing the impact of extreme values on the model. As far as the population size of the city is concerned, except for consumption volume, models trained on small- and medium-sized cities appear more accurate than on big cities at high resolution (Fig. 2*E*). Since spatial heterogeneity in small- and medium-sized cities is less than that found in big cities, suggesting that small/medium cities may be more predictable for socioeconomic characteristics than big cities.

From the application perspective, one important tradeoff is between the number of samples used to train a model and the accuracy that can be reached, because this tradeoff directly determines the cost–benefit of the model application. To calculate this tradeoff relationship, we fix grid size at 3 km and randomly select subsamples of the training set. Results for Beijing, as depicted in Fig. 2*F*, show that even collecting very few random samples can result in fairly high prediction accuracy. Using 10% of the training samples ($\sim$ 100 observations) achieves 86% (SD = 2.3%) accuracy for daytime population, 87% (SD = 2.0%) for nighttime population, 74% (SD = 3.3%) for number of firms, and 82% (SD = 2.6%) for consumption volume. Collecting more samples could increase the accuracy but with diminishing marginal returns (Fig. 2*F*).

To justify the predictive power of restaurants, we include "night-time light"—the most commonly used proxy for urban activities—as baseline models (*SI Appendix*, Table S2). As shown

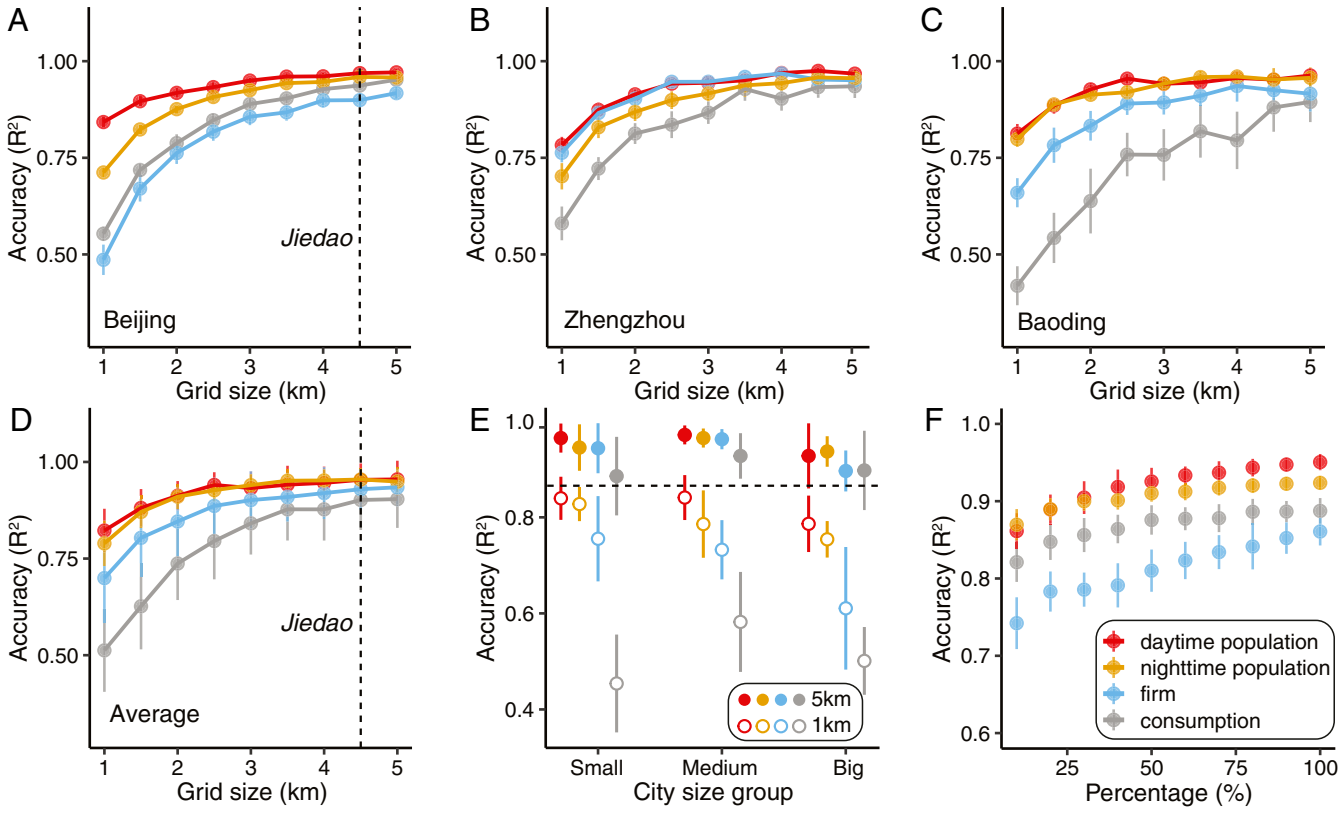

Fig. 2. Prediction accuracy. (*A*) Beijing. (*B*) Zhengzhou. (*C*) Baoding. (*D*) Averaged accuracy of 9 cities. The red, yellow, gray, and blue lines represent the accuracy of daytime population, nighttime population, firms, and consumption, respectively. (*E*) The relationship between city size and model accuracy. Big cities are cities with a population of over 10 million: Beijing, Shenzhen, and Chengdu in our sample. Medium-sized cities have a population of 3 to 10 million: Shenyang, Zhengzhou, and Kunming. Small-sized cities have a population of less than 3 million: Baoding, Yueyang, and Hengyang. The hollow and solid circles represent models trained at 1-km resolution and 5-km resolution, respectively. (*F*) The percentage of training samples and accuracy. Models were trained from the data of Beijing at 3-km resolution.

in *SI Appendix*, Table S2, the "night-time light" performs well in daytime/nighttime population estimation. However, the predictive power of the nighttime lighting is still far from the accuracy of the restaurant model, especially for firm and consumption. In terms of mean squared error (MSE), the predictive error of the restaurant model is only one quarter of that of nighttime lighting.

Heterogeneous Effect. To draw further information from the model, we investigate the heterogeneity of the predicting results across different spatial locations, demographics, and firm industries.

Fig. 3*A* shows the spatial distribution of the predictive error of the daytime population in Beijing, and *SI Appendix*, Fig. S9 gives results for additional cities. Good prediction accuracy is achieved near the urban center. The relatively lower accuracy exists around suburban areas. This differs from remote-sensing data sources, which are more likely to underestimate population in the densely populated areas (23). Night lighting and other satellite data have the saturation effect—making it difficult for the model to accurately predict high-density populations.

Another general pattern in Fig. 3*A* is that the model overestimates the daytime population in some residential areas and underestimates in industrial areas (this pattern reverses itself for the nighttime population model; *SI Appendix*, Fig. S10). This can be explained by the fact that restaurants reflects a "combination" of employment and residents at the neighborhood scale. Thus, the model can achieve good performance in highly mixed land use areas, i.e., urban centers, and lower performance in the opposite direction, i.e., suburban residential or industrial zones. From another angle, we can also treat underestimated regions as potential business opportunities for the restaurant industry.

Without the demographic information available in the mobile phone dataset, we use census data as an alternative to explore predictive accuracy across different groups of people. The latest census in China was administrated in 2010, and the highest spatial resolution data available is at the *Jiedao* level (similar to neighborhoods in the United States), which includes 309 observations in Beijing. Note that there are very few attributes in the census data at the *Jiedao* level, and some important variables like education, religion, and income are not accessible. Here, we include age structure (3 groups: 0 to 14 y, 15 to 64 y, 65+ y) and percentage of immigrants as variables to be predicted. We also take median housing price as a proxy for household wealth. Housing price data are collected from Lianjia.com, the largest real estate agent company in China.

Fig. 3*B* presents the prediction accuracy of different demographic groups. The highest accuracy is achieved for the 15- to 64-y-old group ($R^2 = 0.73$), as they are the main force of urban activities and customers of the restaurant industry. Closely following are the household wealth ($R^2 = 0.70$), percentage of immigrants ($R^2 = 0.59$), age group over 65 y ($R^2 = 0.59$), and age group of below 14 y ($R^2 = 0.56$). Note that there is a 7-y gap between demographic data and restaurant data, and this gap may lower the predictive power of the restaurant model, which could be evaluated with more timely survey data.

Different types of firms have very different preferences for spatial locations. For example, high-skilled firms are more likely to benefit from agglomeration and thus to be concentrated near the city center; low-skill firms, in contrast, are more

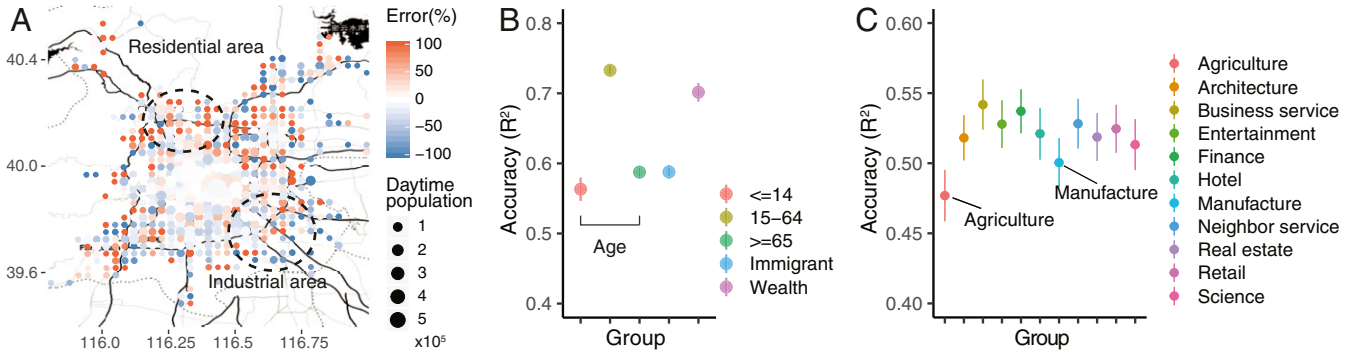

Fig. 3. Heterogeneity. (*A*) Spatial distribution of prediction errors of daytime population at 3-km resolution (Beijing). (*B*) Prediction accuracy of different age groups, immigrant percentage, and wealth (housing price) at the Jiedao level. (*C*) Prediction accuracy of different industries of firms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

sensitive to land price, leading to a dispersed spatial distribution. We train the model by firm category in terms of the industry code and investigate the predictive accuracy across different industries. Fig. 3*C* shows that agriculture and manufacturing are industries with the lowest prediction accuracy. Business service, entertainment, finance, and many other service industries have nearly the same prediction accuracy. This is in line with our hypothesis, as the spatial variation of firms causes an uneven distribution of employees with different income levels and skill sets, which is also reflected in the distribution of restaurants.

SI Appendix, Table S3 shows the top predictors for demographic and firm categories (*Materials and Methods*). For example, 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 to 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 (also noted in ref. 18), other pairs are more elusive; there is no obvious connection between "Bakery" and the presence of immigrants.

Transferable. Finally, based on data from 9 different cities, we study the transferability of the restaurant model—to what extent the model trained with one city's data can estimate other cities' socioeconomic variables. Exploring whether one specific city's model generalizes across regions is helpful. Since most of the "big data"-related research or applications are taken from big cities, if models trained in big cities with rich data sources can be transferred to small cities, then small cities could also benefit from big data and algorithm.

For the cross-city model, we fix the spatial resolution to 3 km and traverse all city pairs. For each pair "A–B," we train the model with any of A's data that shares features with B and apply the model to predict B's outcomes. The results are summarized in Fig. 4. We show that for all variables of interest, as expected, most of the models trained in-city (diagonals) outperform models out-of-city. In some cases, such as firm prediction in Shenyang, the out-of-city models can outperform in-city models (Fig. 4*C*). Generally, models appear to transfer well across cities. This transferability is particularly evident in the daytime/nighttime population and firm datasets—R^2 exceeds 0.7 in most out-of-city models. We also find that models trained in big cities to infer outcomes for other cities outperform models applied in the opposite way in most cases (*SI Appendix*, Table S4). These results indicate that restaurant data can capture common indicators of socioeconomic outcomes, and these commonalities can be transferred by models trained in cities with rich survey data and estimate outcomes of interest with reasonable accuracy in cities where survey outcomes are unobserved.

For the transfer models, we summarize the top predictors in *SI Appendix*, Table S5. "Jiangsu/Zhejiang," "Guangdong," and "Crayfish" appear as top features for all 4 variables of interest, showing the popularity of these cuisines in different cities. "Coffee" only enters the top features in the model for the daytime population. Similar to *SI Appendix*, Table S3, the important predictors learned by the model warrant further exploration.

Discussion and Conclusion

Measuring and mapping the socioeconomic outcomes of cities are of great importance to policy makers and researchers. Here, we demonstrate that local restaurants can accurately infer the spatial distribution of socioeconomic activities within cities at high granularity. Notably, 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. Another potential implication of our work is helping city governors optimize decision-making regarding the efficient allocation of public facilities with the inferred daytime and nighttime population distributions.

The similarity between restaurants and other geolocated digital traces (e.g., online maps) suggests that the potential to reveal a neighborhood's attributes is unlikely to be limited to restaurants. Moreover, the rich attributes of digital traces also make it possible to measure outcomes that were never included in traditional datasets (3).

Taking the national perspective, we show that models trained in one city can achieve good predictive accuracy when applied to other cities. 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. Currently, we only demonstrate the transferability between cities within the same country. One important following question should be whether the restaurant model (or more generally, the socioeconomic predictive model) can be transferred even between different countries? Addressing this issue is beyond the scope of this research, but it is a promising direction for further exploration.

Given the limited availability of high-resolution time series data for population, employment, and other 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. Such investigation should be possible in

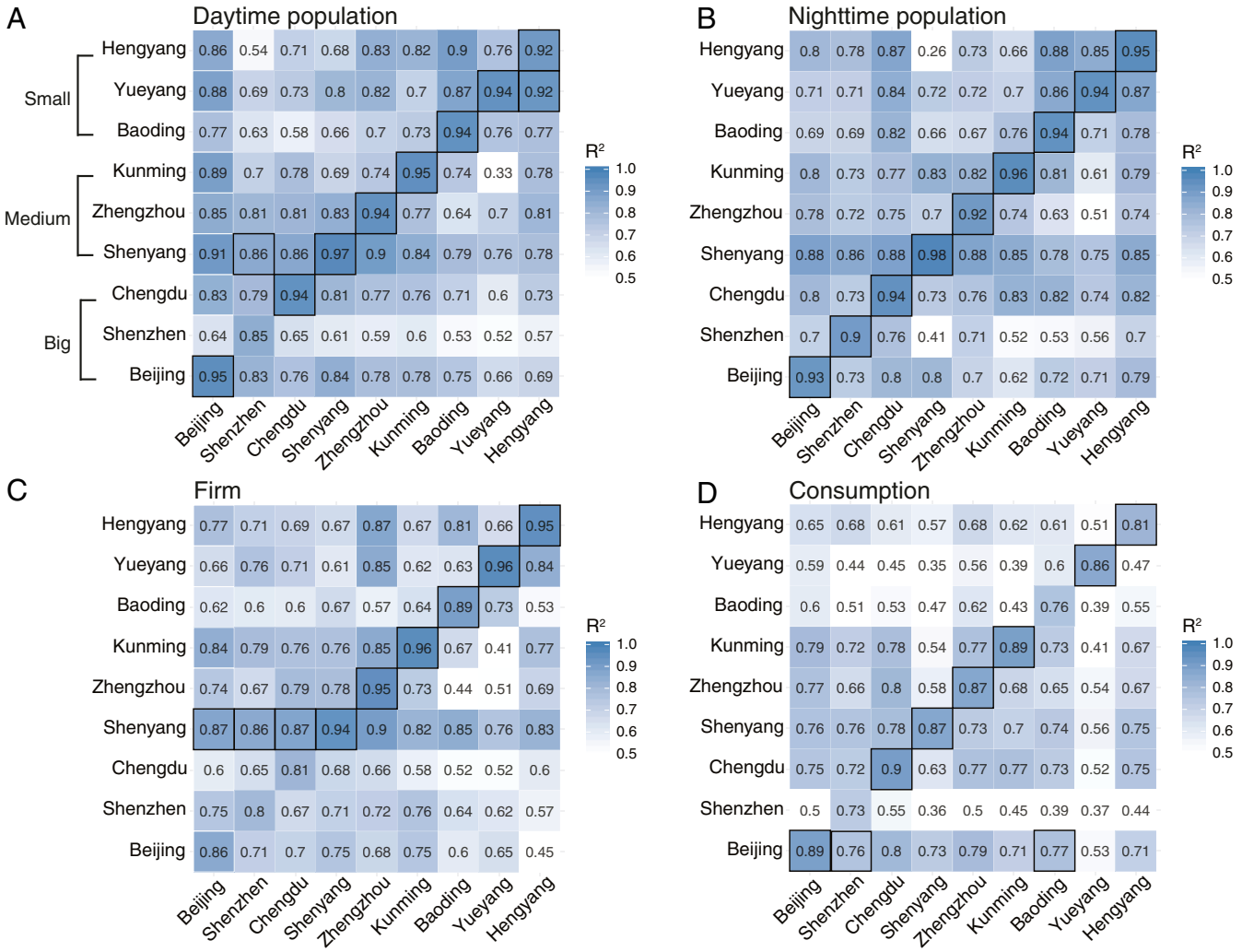

Fig. 4. Cross-city model generalization. Cross-validated R^2 of models trained in one city and applied to other cities. (*A*) Daytime population. (*B*) Nighttime population. (*C*) Firm. (*D*) Consumption. Cities on the *x* axis indicate where the model was trained, with cities on *y* axis showing where the model was evaluated. Reported R^2 values are averaged over 50 trials, and the highest value for each city is within the black box.

the future as survey data and restaurant data are more regularly and frequently gathered. Another untouched but very important direction is deriving high granular data from coarse aggregated sources like census data or other survey instruments. Newly developed algorithms in machine-learning communities, such as superresolution in computer vision, show us the possibility of interpolating aggregated city data with geolocated data sources not limited to restaurants (24).

As emerging areas, big data and machine learning are playing important roles in the urban agenda. However, "there are a number of gaps between making a prediction and making a decision" (25). This is especially important for cities, since most decisions about city development are long-term decisions (related to durable infrastructure and long-term land-use changes), while the current predictive models are trained using short-term data. The approach proposed here could be useful for inexpensively producing granular data on socioeconomic outcomes of interest. However, how to use these data to assist decision-making requires more in-depth research and practice.

Materials and Methods

Restaurant Data. We collected restaurant data in 2017 to 2018 from dianping.com. By the end of 2017, there were 8 million restaurants listed on Dianping, covering 2,298 (80%) county-level administrations in China. It should be noted that although Dianping data cover a very large share of the restaurant industry, the penetration rate still does not reach 100%. Using Beijing as an example: according to the report by Beijing Cuisine Association, as of the end of 2016, there were 147,575 restaurants in operation. In our Dianping restaurant data, we have 139,131 restaurants, which accounts for 94% of the total number of restaurants.

For the other cities, we have collected 98,531 (Shenzhen), 134,497 (Chengdu), 57,915 (Shenyang), 66,458 (Zhengzhou), 56,140 (Kunming), 46,219 (Baoding), 15,864 (Yueyang), and 18,209 (Hengyang) restaurants, respectively. In total, we have collected 632,964 restaurants for 9 cities.

Daytime and Nighttime Population Data. The distributions of daytime and nighttime populations were estimated by mobile phone location data for the year 2015. The raw data are generated when people use location-based services; it contains a series of geopositioning points (time stamp, longitude, latitude) for each anonymous user. To infer the daytime and nighttime population distributions, we take following steps (see *SI Appendix*, Note 1 for details). 1) We identify each user's stay points, defined by when moving distance is less than 200 m within a 10-min time threshold. 2) We then cluster the stay points into different clusters using density-based spatial clustering of applications with noise (DBSCAN) algorithm. 3) Finally, we apply Xgboost, a tree-based machine-learning algorithm, to train 2 classifiers for the home and work location classification, respectively. The distributions of home and work locations are regarded as the nighttime and daytime population distributions, respectively. After data preprocess, we have 56.3 million (mobile phone inferred) population for 9 cities (see *SI Appendix*, Table S1

for details). To protect user privacy, the datasets are anonymized during the whole process and aggregated into 100 m × 100 m grid cells for further analysis.

At the aggregated level, we calculate the correlation between mobile phone-inferred home locations (nighttime distribution) and the population microcensus data at the district level. The R^2 is 0.97 for Beijing, indicating that the mobile phone inferred population fits well with the microcensus data in terms of spatial distribution (*SI Appendix*, Fig. S11).

Firm Data. We collected firm registration record data (2000 to 2017) from the registry database of the State Administration for Industry and Commercial Bureau of China. These data cover the registered information for all firms in China, with variables including firm name, year established, operation status, address, industry code, and so on. There are 8.72 million firms in these cities we study (*SI Appendix*, Table S1). We geocode firm addresses into longitude and latitude using Baidu Map application program interface and then aggregate firms by grid cells of each city.

Consumption Data. Consumption data were aggregated by the bank card records for point-of-sales (POS) from July to September in 2016. The original record is anonymized, and we only know the amount of consumption in each location. These data have several limitations. First, the consumption category (e.g., food, hotel, transportation, etc.) is not included in this dataset, making it impossible to distinguish food consumption from all records. Second, POS and bank cards have different adoption rates across cities, which may affect the model's generalizability. However, it is the highest level of spatial granularity dataset we could access to estimate the consumption at grid cell level across different cities. To reduce the noise caused by outliers, we set an upper limit for each record—any single purchase of more than 10,000 renminbi (RMB) was set to the upper limit (10,000 RMB).

Prediction Model. To estimate the grid cell level socioeconomic outcomes, we train LASSO regression models using the *glmnet* package (26) in R. As a commonly used machine-learning method, LASSO performs both regularization and variable selection by introducing ℓ_1 penalty, which minimizes:

$$\min_{\beta_0,\boldsymbol{\beta}} J(\beta_0,\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \beta_0\mathbf{1} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \quad [1]$$

where $\lambda(\geq 0)$ controls the sparsity of the model and could be selected using cross-validation. By shrinking the coefficient of some variables to zero, LASSO effectively reduces overfitting and improves prediction accuracy on small datasets. To avoid overfitting, we adopt the following process. First, we randomly divide the data into training (80%) and test (20%) sets. For the training set, we train the model using fivefold cross-validation (splitting the training set into 5 randomly sampled folds; 4 folds are used to train the model and then test the accuracy on the fifth fold). This procedure is repeated 50 times to determine the average accuracy on the test set.

Variables' Importance. Although LASSO can shrink many coefficients to zero and report the value of "used" coefficients, we usually cannot directly compare coefficients to show variables' contribution. As demonstrated in ref. 27, a variable used in one partition of the dataset may not be used in another (although the prediction accuracy of different partitions is similar). We assume that important variables should appear more times than unimportant ones in different partitions. Thus, we compare the frequency with which different variables appear over 50 trials.

Data Availability. All data and code needed to replicate this research can be downloaded from https://github.com/leiii/restaurant (28).

ACKNOWLEDGMENTS. We thank Hao Li and Meng Li for their great assistance with data preprocessing and Kevin P. O'Keeffe, Hongbin Pei, Zhan Shi, Haishan Wu, and seminar participants at The Institute for Operations Research and the Management Sciences, Massachusetts Institute of Technology (MIT) for helpful comments. We also thank Allianz, Amsterdam Institute for Advanced Metropolitan Solutions, Brose, Cisco, Ericsson, Fraunhofer Institute, Liberty Mutual Institute, Kuwait-MIT Center for Natural Resources and the Environment, Shenzhen, Singapore-MIT Alliance for Research and Technology, Uber, the Vitoria State Government, Volkswagen Group of America, and all of the members of the MIT Senseable City Lab Consortium and the China Future City Lab for supporting this research. This work was partially supported by National Natural Science Foundation of China Grants 41801299 and 71625004.

1. N. Wardrop *et al.*, Spatially disaggregated population estimates in the absence of national population and housing census data. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3529–3537 (2018).
2. L. Einav, J. Levin, Economics in the age of big data. *Science* **346**, 1243089 (2014).
3. E. L. Glaeser, S. D. Kominers, M. Luca, N. Naik, Big data and big cities: The promises and limitations of improved measures of urban life. *Econ. Inq.* **56**, 114–137 (2018).
4. X. Chen, W. D. Nordhaus, Using luminosity data as a proxy for economic statistics. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 8589–8594 (2011).
5. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring economic growth from outer space. *Am. Econ. Rev.* **102**, 994–1028 (2012).
6. D. Donaldson, A. Storeygard, The view from above: Applications of satellite data in economics. *J. Econ. Perspect.* **30**, 171–198 (2016).
7. M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5802–5805 (2013).
8. H. Choi, H. Varian, Predicting the present with google trends. *Econ. Rec.* **88**, 2–9 (2012).
9. J. L. Toole *et al.*, Tracking employment shocks using mobile phone data. *J. R. Soc. Interf.* **12**, 20150185 (2015).
10. L. Dong *et al.*, Measuring economic activity in China with mobile big data. *EPJ Data Sci.* **6**, 29 (2017).
11. P. Deville *et al.*, Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15888–15893 (2014).
12. J. Blumenstock, G. Cadamuro, R. On, Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
13. A. Cavallo, Scraped data and sticky prices. *Rev. Econ. Stat.* **100**, 105–119 (2018).
14. N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, C. A. Hidalgo, Computer vision uncovers predictors of physical urban change. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 7571–7576 (2017).
15. N. Jean *et al.*, Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
16. T. Gebru *et al.*, Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 13108–13113 (2017).
17. E. L. Glaeser, H. Kim, M. Luca, "Nowcasting the local economy: Using yelp data to measure economic activity" (NBER Working Paper 24010, National Bureau of Economic Research, Cambridge, MA, 2017).
18. E. L. Glaeser, H. Kim, M. Luca, Measuring gentrification: Using yelp data to quantify neighborhood change. *AEA Pap. Proc.* **108**, 77–82 (2018).
19. D. R. Davis, J. I. Dingel, J. Monras, E. Morales, How segregated is urban consumption?*J. Polit. Econ.*, in press (2019).
20. J. S. Kang, P. Kuznetsova, M. Luca, Y. Choi, "Where not to eat? Improving public policy by predicting hygiene inspections using online reviews" in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Seattle, WA, 2013), pp. 1443–1448.
21. R. Tibshirani, Regression shrinkage and selection via the LASSO. *J. Roy. Stat. Soc. B Met.* **58**, 267–288 (1996).
22. D. Martin, S. Cockings, S. Leung, Developing a flexible framework for spatiotemporal population modeling. *Ann. Am. Assoc. Geogr.* **105**, 754–772 (2015).
23. A. E. Gaughan *et al.*, Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Sci. Data* **3**, 160005 (2016).
24. T. Vandal *et al.*, "DeepSD: Generating high resolution climate change projections through single image super-resolution" in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computer Machinery, New York, NY, 2017), pp. 1663–1672.
25. S. Athey, Beyond prediction: Using big data for policy problems. *Science* **355**, 483–485 (2017).
26. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1 (2010).
27. S. Mullainathan, J. Spiess, Machine learning: An applied econometric approach. *J. Econ. Perspect.* **31**, 87–106 (2017).
28. L. Dong, C. Ratti, S. Zheng, Predicting neighborhoods' socioeconomic attributes using restaurant data. GitHub. https://github.com/leiii/restaurant. Deposited 3 May 2019.