SUPPLEMENTARY INFORMATION

Supplementary Information for "Economic Outcomes Predicted By Diversity in Cities"

Shi Kai Chong[1†], Mohsen Bahrami[2,3*†], Hao Chen[4], Selim Balcisoy[5], Burcin Bozkaya[3,6] and Alex 'Sandy' Pentland[1]

[*]Correspondence:
bahrami@mit.edu
[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Full list of author information is available at the end of the article
[†]Equal contributor

Datasets, basic statistics, and analysis results

Here we provide detailed information about the datasets used for this research. We used various datasets to study urban environment in three different countries, namely: Turkey, China, and the United States. In the following sections we provide detailed information and basic statistics of datasets utilized for analyses. We also provide further details about the analysis results that are not provided in the manuscript. Finally, for replication purposes, we made all required data and code available at: github.com/cshikai/Cities

1 Istanbul Datasets and Analysis Results

1.1 Istanbul Districts

Istanbul is a metropolitan with a population size of more than 15 million. It is divided into 39 districts. Districts are administrative areas within the city and each district has a local government and municipality. Many census statistics are provided at districts level. Figure 1, shows a satellite image of Istanbul provided by Google Maps. Istanbul districts shape files for reproduction purposes are available at: github.com/cshikai/cities/tree/master/data/istanbul/district_level_shape

Figure 1 Satellite image of Istanbul provided by Google Maps

Table 1, shows basic statistics about districts population, area, and population density. Figure 2, shows Istanbul choropleth map indicating the population size per district.

Figure 3, provides bar plots of Istanbul's districts population, area, and population density. Numeric data for population and districts area are obtained through the

Table 1 Basic statistic of districts population, area (km^2), and population density ($population/km^2$)

Istanbul Districts	Min	Mean	Median	Max
Population	15,623	375,832	354,882	761,064
Area	7.2	137	38	1,040
Population density	43	13,715	10,533	42,977

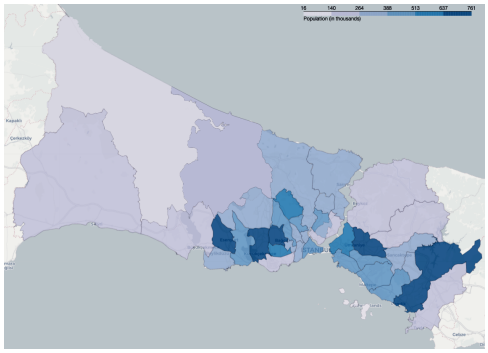

Figure 2 Istanbul districts colored by population volume

census by the Turkish Statistical Institute in Turkey (www.turkstat.gov.tr) and made available at: github.com/cshikai/cities/tree/master/data/istanbul

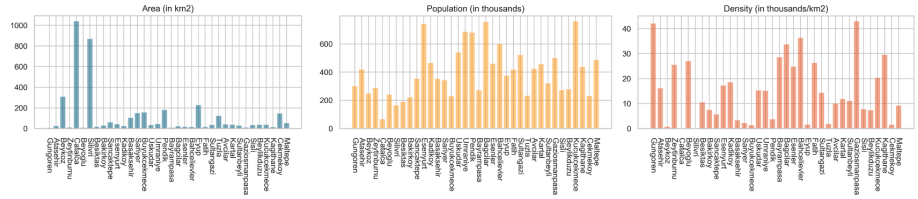

Figure 3 Bar plots of Istanbul districts area (km^2), population ($thousands$), and population density ($thousands/km^2$)

1.2 Housing prices

Rent data was provided by the leading real estate listing company in Turkey (www.hurriyetemlak.com) for research purposes . The data contains monthly and yearly basic rent statistics at Istanbul districts level from 2013 to 2016. The statistics are maximum, minimum, median, and average rent per squared meter. Figure 4, shows the rent amount distribution over Istanbul districts as of 2016. Complete data for rent per m^2 at districts level, 2013 to 2016 is available at: github.com/cshikai/cities/blob/master/data/istanbul/housing_price.csv

1.3 Points of Interest Data

This dataset is provided by Here.com which is a commercial map production company. The company collects data via data collection vehicles moving around in every region. They also use several external dataset provided for them by local organizations to enrich their maps. Since the dataset is not crowd sourced, we contend that the level of bias is low in the maps created by this company. We used POI

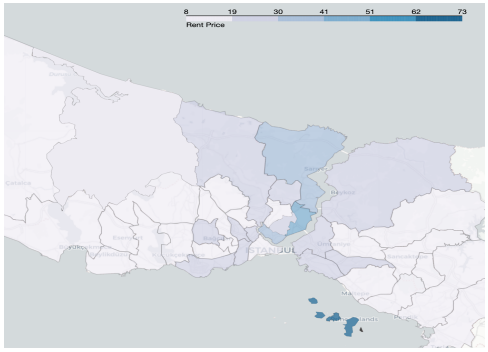

Figure 4 Istanbul districts colored by yearly rent per m^2

data and maps for years 2015 and 2016, updated quarterly (every three months). In the dataset, POIs are grouped into twelve types. POI types with examples are as below:

- Community service centers (such as local government offices, law courts, post offices, religious places, and wedding venues)
- Financial institutes (such as banks, exchange places, lending and financing offices)
- Educational institutes (including schools, universities, and private education centers)
- Business Centers (such as commercial, managerial, and other types of business offices)
- Entertainment places (such as Bars, CAFE, Disco, places for computer games, and art galleries)
- Shopping places (such as supermarkets, grocery stores, clothing, and corner shops)
- Restaurants (all types of eating places such as fast food, Pizza, and traditional food)
- Hospitals (all places which provide health services such as hospitals, clinics, dentists, etc.)
- Parks (including public parks, public and private sport places, and gyms)
- Travel destinations (all tourism and lodging related places)
- Auto services (any type of place that provides any type of service for vehicles such as gas stations, car wash, and maintenance services)
- Transportation hubs (such as subway and railway stations, terminals, and passenger used sea ports)

Figure 5, shows the POI count and related diversity distributions at each district. Bar plots for each single POI type are available at:
github.com/cshikai/cities/tree/master/huffmodel_poi_testing/plots
Table including count of each POI type at each district is downloadable at: github.com/cshikai/cities/blob/master/data/istanbul/attractiveness.csv

1.4 Credit Card Data

This dataset contains 4,254,652 geo-tagged credit card transactions in Istanbul that covers the expenditure of 62,392 customers from July 1, 2014 to June 30, 2015.

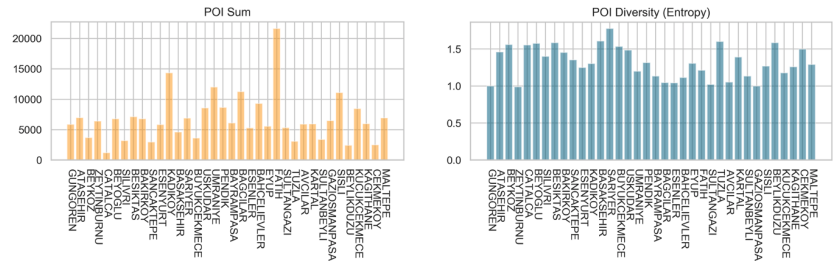

Figure 5 Bar plots of total POI counts (left) and POI diversity (right) at Istanbul districts

Transactions made in 75,448 unique merchants. The transaction records contain hashed customer IDs, transaction amounts, merchants' business categories, and their locations. Customer information dataset includes customers' demographic information such as their age, gender, marital status, job type, education level, income, and their home and work location. Table 2 shows the number of transactions in top four merchant categories. Table 2 indicates that the customers tend to visit grocery stores more often than other categories.

Table 2 Number of transaction in top four most frequently visited merchant categories

Merchant category	Transaction count
Grocery Stores	801,818
Gas stations	422,858
Clothing stores	377,138
Restaurants	39,539

Figure 6, shows the population distribution heat map provided by ESRI (left), and the heat map of the sampled bank customers (right). Esri is an international supplier of GIS software (www.esri.com). Despite the difference in map tiles, resolution, and color pallets used to create the maps, the distribution similarity is clear.

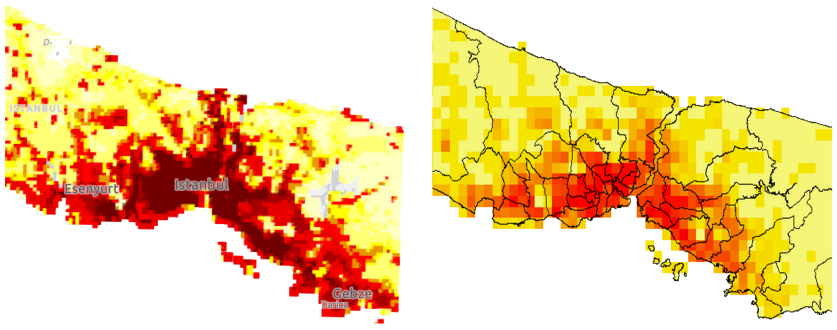

Figure 6 Heat map of population distribution (left) vs heat map of the sampled bank customers' home locations (right)

The number of customers per district shows a 0.817 correlation with the districts' population during the time frame data was collected. Figure 7, shows the scatter plot of number of sampled customers vs the population at district level. In addition, there is a 0.6115 correlation between the average yearly spending amount per bank

customers and average yearly household income in districts. Similar type of correlation coefficient analysis was previously used by Di Clemente et al. [1] to validate the representatives of credit card data. Figure 8 shows the scatter plot of average yearly spending amount per bank customers and average yearly household income in thousands in Istanbul districts. It is important to note that the bank that provided the datasets, is one of the leading banks in Turkey, which had more than 11% market share of active credit cards 2014-2016. Thus, we contend that the sample is well balanced across the metropolitan of Istanbul and representative. Istanbul districts income level data is available at: www.mahallemistanbul.com

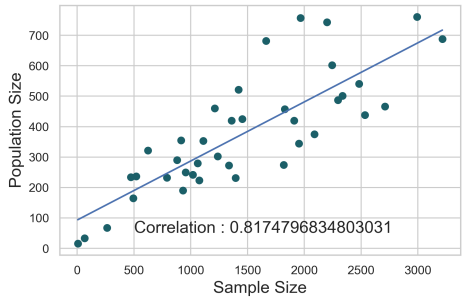

Figure 7 Sampled bank customer count vs population size at each district

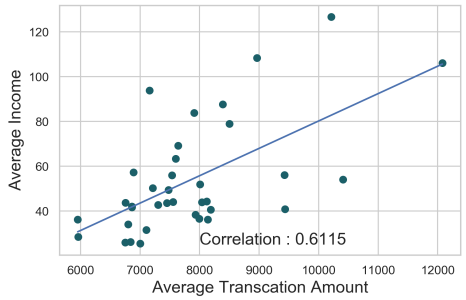

Figure 8 Average yearly spending amount per bank customers vs average yearly household income in thousands in Istanbul districts

All required information from the credit card dataset for replication of this study is available in various tables at:
github.com/cshikai/cities/tree/master/data/istanbul
The tables include customer demographics, merchants business type and related category code, customer set of each merchant, and all required inter-district flow matrices.

1.5 Economic Growth - GDP proxy

We measure economic growth as the percentage change in GDP. As mentioned in the manuscript, the private insurance company that provided us the data is one of the leading insurance companies in Turkey. The dataset we used, contains records for

around 3.8 million customers with more than 180 million transactions in all eleven insurance categories provided by the company during three years of 2014-2016. The insurance types are: Home/House insurance (complete coverage), workplace/Office insurance (complete coverage), Vehicle Liability and Collision insurances, Individual and group health care plans, House/Workplace Earthquake insurance, Fire damages insurance, Engineering insurance, Transportation insurance, Responsibility insurance, and Individual accident and life insurance. As a proxy for GDP we considered four first types (above), and it is important to note that the insurance for those four categories are all contracted by individuals. Thus, we contend that the numbers could be a reasonable proxy for economic activity. Moreover, a very high correlation coefficient in the city level supports our belief. Numbers are downloadable via the following link: github.com/cshikai/cities/blob/master/data/istanbul/gdp.csv . Figure 9, shows the map of Istanbul districts colored by their economic productivity for the year 2016.

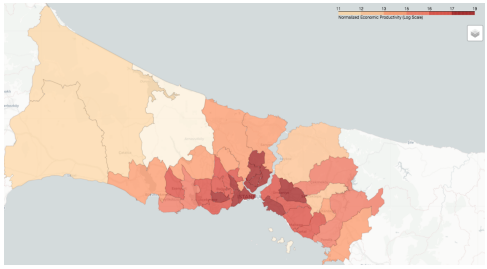

Figure 9 Istanbul districts productivity as of 2016.

1.6 Flow modelling and parameter fitting

We utilized three different methods for fitting the parameters α, β, γ in the flow model. The first two involve fitting a generalized linear model (GLM) with Poisson and Negative Binomial distributions on the count of flows [2]. The third method involves normalizing the flow counts by the total count of each district i to obtain the probability of movement to j given that the origin is i. We then linearize the data and fitted the parameters via ordinary least squares regression according to the methodology listed in Huff & McCALLUM (2008) [3] available at: www.esri.com/library/whitepapers/pdfs/calibrating-huff-model.pdf

The goodness-of-fit of each model is shown in table 3. We found that the third method (via OLS) has the best result, as even though the Poisson model has the best fit according to the Pseudo R2 value, the large residual deviance suggests that it is a poor choice of model. Residual plots of the models are shown in figure 10.

Table 3 Goodness-of-fit of flow models and optimal parameters

Model	Root Mean Squared Error	Pseudo R2	Residual Deviance	Df Residual	Intercept	Alpha	Beta	Gamma
Poisson	3141.931822	0.800381	1.478851e+06	1292.0	6.455943e+00	0.765199	2.393633	-1.681224
Negative Binomial	3765.940072	0.599774	1.833065e+03	1292.0	4.352270e+00	0.817191	4.041559	-1.432145
Gaussian	2472.310230	0.622938	1.452620e+03	1291.0	4.510281e-17	0.998249	3.449586	-1.958913

In addition to fitting a single set of parameters to the flows of all the districts, we also independently fitted sets of parameters to flows out of each district, and the results are shown in Figure 11.

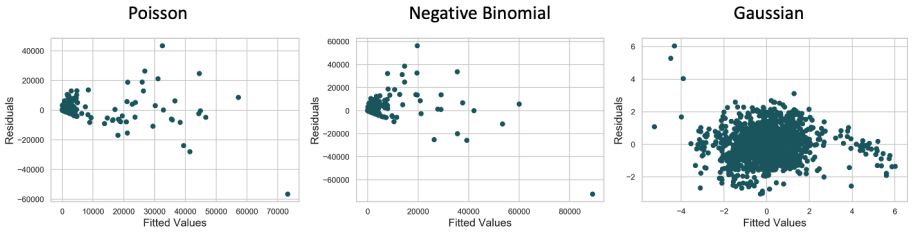

Figure 10 Residual Plots for Flow Models

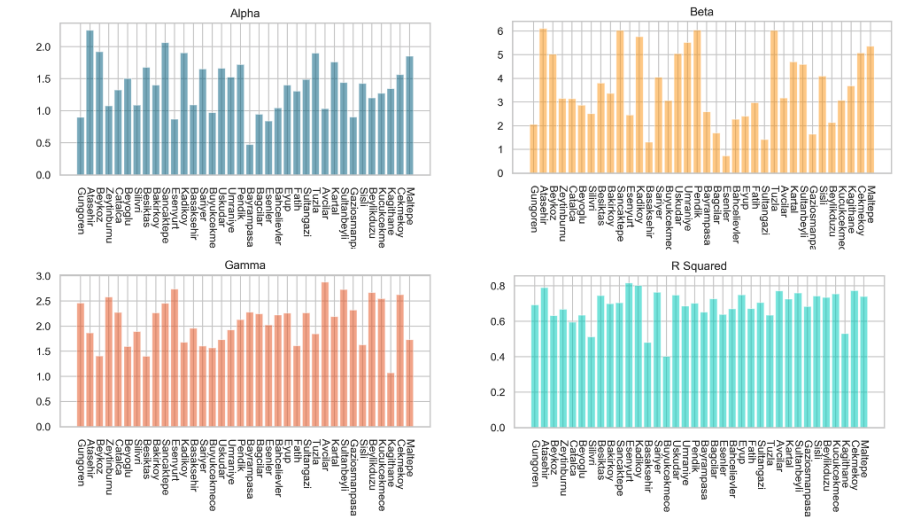

Figure 11 Parameters and R^2 values of districts' Huff Models after independent parameter optimization

1.7 Regression Analysis

For all the regression analyses we considered different transformation of the variables, and the final model was chosen based on the highest resulting R^2. For the case of modelling economic productivity using flows, we considered both linear and log transformation of the variable. We chose and provided the linear model results in the manuscript as it was producing slightly better R^2. Figure 12, shows plots for both linear and log transformation models. Table 4 shows the regression results of predicting growth using consumption diversity, both with and without control variables.

2 Beijing, China Datasets and Analysis Results

2.1 Beijing Districts

Beijing is split into 16 administrative area, which is further divided into 286 regions at the township level. We refer to each of these regions as a 'district' in this paper. For our study, data is obtained on 187 of these districts, and districts that are not considered in this study (due to the lack of available data) are depicted in black in Figure 14. Figure 13, shows a satellite image of Beijing provided by Google Maps.

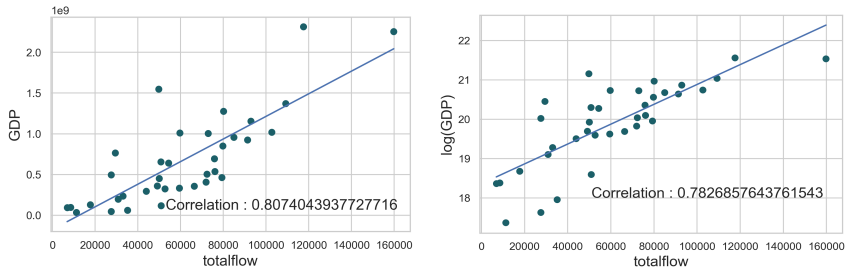

Figure 12 Scatter plot of linear (left) vs log (right) transformation models

Table 4 Regression coefficients for prediction of Economic Growth using Consumption Diversity (Istanbul)

	Model 1	Model 2	Model 3	Model 4
$ConsumptionDiversity, H$	125.4761*** (21.3132)	125.5035*** (20.5040)	126.2242*** (21.4657)	124.7827*** (21.3772)
$PopulationDensity, \rho$		-0.5636* (0.2916)	-0.5588* (0.2980)	-0.9779** (0.4645)
$HousingIndex, I$			-0.4669 (3.3957)	-3.1322 (4.0709)
$GeographicCentrality, D$				115.0701 (98.2102)
$Constant$	-225.9388*** (38.3526)	-217.6448*** (37.1451)	-217.9009*** (37.7559)	-220.2876*** (37.5931)
$Adj.R-squared$	0.4902	0.5282	0.5138	0.5194
$R-squared$	0.5048	0.5552	0.5554	0.5743
$Observations$	36	36	36	36

Beijing districts shape files for reproduction purposes are available at: github.com/cshikai/cities/tree/master/data/beijing/bj_shapefile

Figure 13 Satellite image of Beijing provided by Google Maps

Table 5, shows basic statistics about districts population, area, and population density. According to tables 1 and 5, the values for size, population, and population density are in the same order of magnitude in both Istanbul and Beijing metropolitan areas.

Table 5 Basic statistic of districts population, area (km^2), and population density ($population/km^2$)

Beijing Districts	Min	Mean	Median	Max
Population	2,391	74,980	51,561	633,107
Area	0.7	59	33	378
Population density	18	9,522	1,895	77,172

Figure 14, shows a Beijing districts choropleth map indicating the population size. Figure 15 shows bar plots of Beijing's district population, area, and population density. Numeric data for population and districts area are obtained through the census, and is available at: github.com/cshikai/cities/tree/master/data/beijing

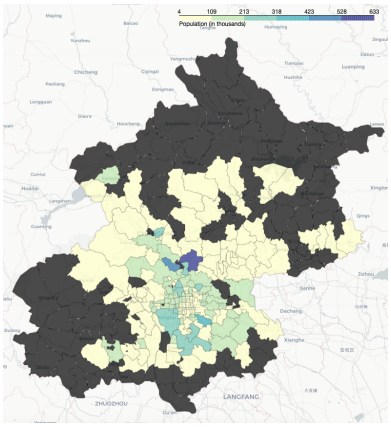

Figure 14 Beijing districts colored by population volume

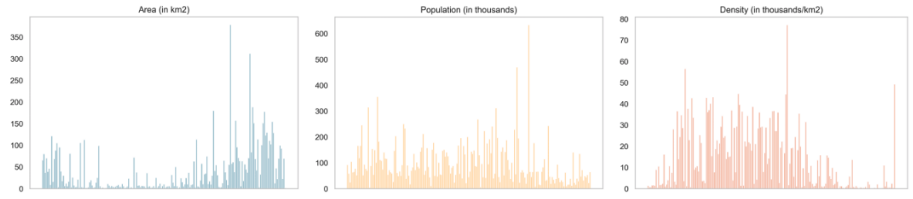

Figure 15 Bar plots of Beijing districts area (km^2), population ($thousands$), and population density ($thousands/km^2$)

2.2 Housing Prices

Real estate prices was obtained from the census. The data contains the mean real estate price per m^2 in 2016 (measured in Chinese RMB). Figure 16, shows the price distribution over Beijing districts as of 2016. Complete data for housing price per m^2 at districts level, is available at:
github.com/cshikai/cities/blob/master/data/beijing/bj_housing_price.csv

2.3 Consumption Data

We analyze data collected from Meituan.com and Dianping.com, which are Chinese group buying websites (i.e similar to combination of groupon and yelp) that are

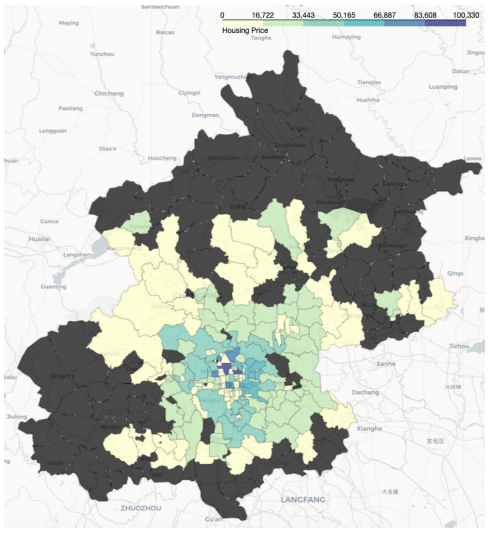

Figure 16 Beijing districts colored by housing price per m^2

extensively used in the community. The dataset consists of 208,360 transactions from 164,170 unique customers and 6,521 unique businesses during four months. For each business, we have information about their location, average rating, category, monthly sales revenue, the products they have offered, and the list of reviewers (for each of their products). Figure 17 shows the distribution of categories of the transactions, and Figure 18 shows the geographical distribution of these transactions across the city.

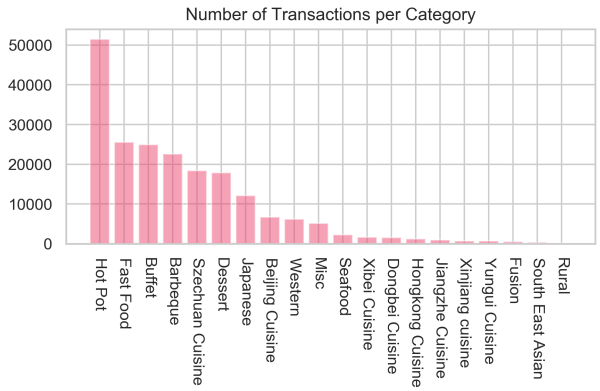

Figure 17 Bar plot of categories of transactions in the Meituan-Dianping Dataset)

2.4 Economic Growth

We measure the economic development in different regions of Beijing by the total capital asset of secondary and tertiary sectors in each Beijing district. Capital accumulation is a key determinant of positive economic growth in many established economic models, and thus we take this metric as an indicator

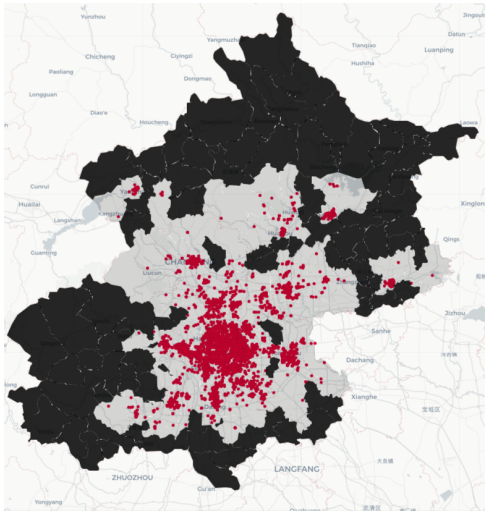

Figure 18 Geographic distribution of transactions in the Meituan-Dianping Dataset

of future economic growth. The date can be accessed via the following link: github.com/cshikai/cities/blob/master/data/Beijing/beijing_econ.csv
Figure 19 shows a choropleth map of Beijing districts' economic productivity for the year 2016.

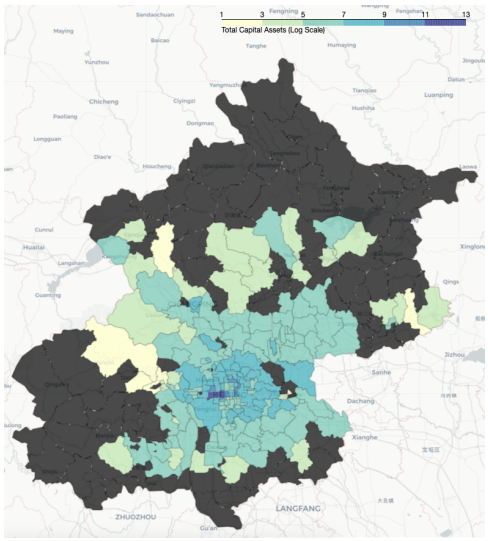

Figure 19 Beijing districts' economic productivity as of 2016

2.5 Regression Analysis

Table 6 shows the regression results of predicting growth using consumption diversity, both with and without control variables.

Table 6 Regression coefficients for prediction of Economic Growth using Consumption Diversity (Beijing)

	Model 1	Model 2	Model 3	Model 4
$ConsumptionDiversity, H$	2.0412*** (0.2333)	1.1986*** (0.1850)	0.8990*** (0.1916)	0.8795*** (0.1861)
$PopulationDensity,\ \rho$		0.2006*** (0.0161)	0.2073*** (0.0155)	0.1555*** (0.0211)
$HousingIndex, I$			0.1199*** (0.0290)	0.1001*** (0.0287)
$GeographicCentrality, D$				10.3531*** (2.9624)
$Constant$	-0.1551 (0.4246)	-0.3217 (0.3138)	-0.4483 (0.3025)	-0.3510 (0.2949)
$Adj.R-squared$	0.2889	0.6123	0.6435	0.6641
$R-squared$	0.2927	0.6164	0.6493	0.6713
$Observations$	187	187	187	187

3 United States Datasets and Analysis Results

3.1 USA Districts

For USA, we studied 29 metropolitan areas, which referred to as 'districts' in this paper. While there are less regions studied in the US dataset as shown in Figure 20, data was available across multiple years from 2011 to 2015 for each district. With the unit of analysis being growth per district per year, we are able to obtain a total of 145 data points representing 29 regions over 5 years. Districts that are not considered in this study (due to the lack of available data) are depicted in black in Figure 20. Bar plots of the districts population, area, and population density distributions are shown by Figure 21. USA districts shape files for reproduction purposes are available at:
github.com/cshikai/cities/tree/master/data/usa/shapefiles

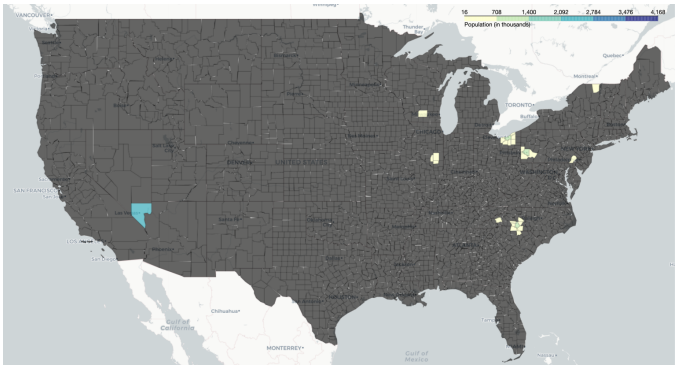

Figure 20 USA districts available in the dataset, colored by population volume

Table 7, shows basic statistics about districts population, area, and population density of the districts available in the USA dataset.

Table 7 Basic statistic of districts population, area (km^2), and population density ($population/km^2$)

USA Districts	Min	Mean	Median	Max
Population	16,387	514,289	213,442	4,167,947
Area	796	3,600	1,710	23,893
Population density	15	177	128	727

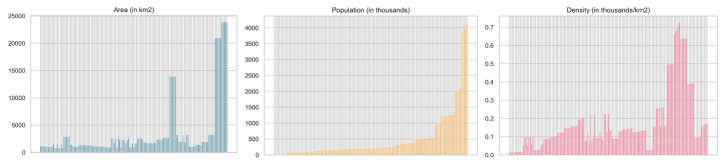

Figure 21 Bar plots of USA districts area (km^2), population ($thousands$), and population density

3.2 Consumption Data

We used data available from the Yelp Data set Challenge to measure consumption in the US areas. The data set consists of 80,326 transactions of 23,150 unique businesses across 5 years from 2011 to 2015.

For each business, we have information about their location, average rating, category. Figure 22 shows the distribution of categories of the transactions, and Figure 23 illusterates the geographical distribution of these transactions across the areas.

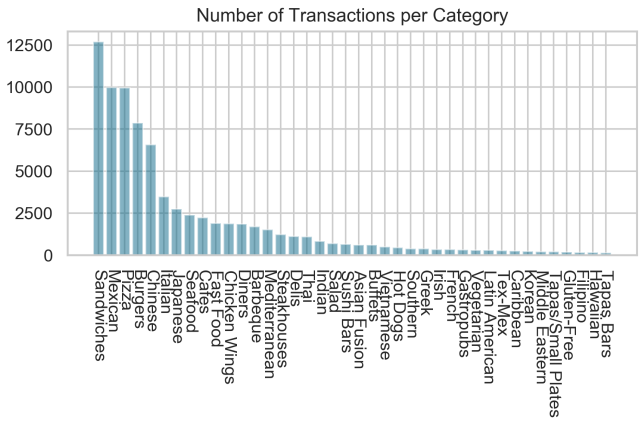

Figure 22 Bar plot of categories of transactions in the Yelp Dataset

3.3 Economic Growth

We approximated economic productivity via the sum of personal incomes in each district, which is obtained via the United States census bureau. The date can be accessed via the following link:

github.com/cshikai/cities/blob/master/data/usa/CensusData

Figure 24 shows a choropleth map of the economic productivity in the areas studied for the year 2015.

3.4 Regression Analysis

Table 8 shows the regression results of predicting growth using consumption diversity, both with and without control variables. As shown in the table, the diversity of consumption is significant in all the models.

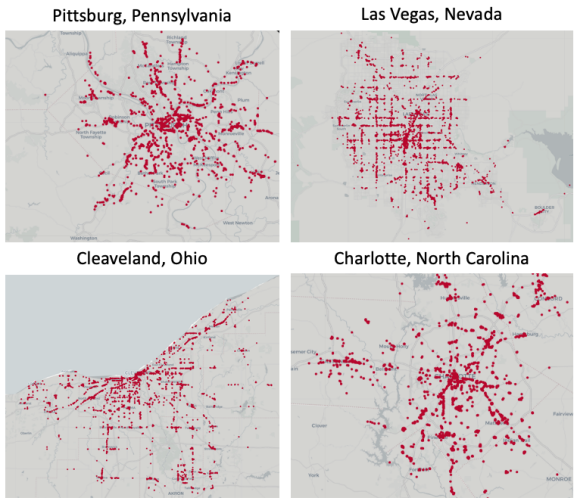

Figure 23 Geographic distribution of transactions in the Yelp Dataset

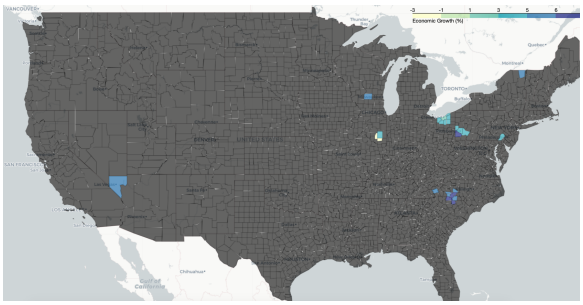

Figure 24 USA districts' economic productivity as of 2015

Table 8 Regression coefficients for prediction of Economic Growth using Consumption Diversity (USA)

	Model 1	Model 2	Model 3	Model 4
$ConsumptionDiversity, H$	0.0232*** (0.0032)	0.0248*** (0.0032)	0.0249*** (0.0033)	0.0258*** (0.0031)
$PopulationDensity, \rho$		-0.0039* (0.0020)	-0.0039* (0.0021)	-0.0059*** (0.0020)
$HousingIndex, I$			0.0053 (0.0123)	0.0170 (0.0122)
$GeographicCentrality, D$				0.0721*** (0.0194)
$Constant$	-0.0087 (0.0073)	-0.0077 (0.0073)	-0.0140 (0.0165)	-0.0374** (0.0170)
$Adj.R-squared$	0.2686	0.2822	0.2781	0.3382
$R-squared$	0.2737	0.2922	0.2931	0.3565
$Observations$	145	145	145	145

Author details
[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]MIT Connection Science, Institute for Data, Systems Society, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]School of Management, Sabanci University, 34956 Istanbul, Turkey. [4]School of Economics and Resource Management, Beijing Normal University, Beijing, China. [5]Faculty of Engineering and Natural Sciences, Sabanci University, 34956 Istanbul, Turkey. [6] New College of Florida, 5800 Bay Shore Road, Sarasota, FL 34243, USA.

References

1. Di Clemente, R., Luengo-Oroz, M., Travizano, M., Xu, S., Vaitla, B., González, M.C.: Sequences of purchases in

credit card data reveal lifestyles in urban populations. Nature communications **9(1)**, 1–8 (2018)
2. Beiró, M.G., Bravo, L., Caro, D., Cattuto, C., Ferres, L., Graells-Garrido, E.: Shopping mall attraction and social mixing at a city scale. EPJ Data Science **7(1)**, 28 (2018)
3. Huff, D., McCALLUM, B.M.: Calibrating the huff model using arcgis business analyst. ESRI White Paper (2008)