to appear: EPJ Data Science, 2020. Comments to pentland@mit.edu

# Economic Outcomes Predicted by Diversity in Cities

**Shi Kai Chong[1,*,+], Mohsen Bahrami[1,2,+], Hao Chen[3], Selim Balcisoy[4], Burcin Bozkaya[2], and Alex Pentland[1]**

[1]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139
[2]School of Management, Sabanci University, Istanbul, Turkey
[3]School of Economics and Resource Management, Beijing Normal University, China
[4]Faculty of Engineering  Natural Sciences, Sabanci University, Istanbul, Turkey
[*]cshikai@mit.edu
[+]these authors contributed equally to this work

## ABSTRACT

Much recent work has illuminated the growth, innovation, and prosperity of entire cities, but there is relatively less evidence concerning the growth and prosperity of individual neighborhoods. In this paper we show that diversity of amenities within a city neighborhood, computed from openly available points of interest on digital maps, accurately predicts human mobility ("flows") between city neighborhoods and that these flows accurately predict neighborhood economic productivity. Additionally, the diversity of consumption behaviour or the diversity of flows together with geographic centrality and population density accurately predicts neighborhood economic growth, even after controlling for standard factors such as population, etc. We develop our models using geo-located purchase data from Istanbul, and then validate the relationships using openly available data from Beijing and several U.S. cities. Our results suggest that the diversity of goods and services within a city neighborhood is the largest single factor driving both human mobility and economic growth.

## Introduction

Cities are known as engines of industry and innovation, but the causal processes that produce these results are complex and the causality is unclear. Moreover, while substantial evidence is accumulating concerning the growth and prosperity of cities as a whole, there is less clarity about the factors and processes that determine the prosperity and attractiveness of individual neighborhoods or districts.

This paper provides a simple and practical method for forecasting local neighborhood prosperity that accounts for around half of the variance in economic growth as well as accurately predicting interaction patterns between neighborhoods (e.g., number of workers and shoppers from other areas of the city). The method has been validated on data from three continents, and sheds new light on the causal processes underlying the evolution of city neighborhoods.

We begin by utilizing open sourced map data and geo-tagged expenditure records from the banking sector, covering a significant portion of the population of Istanbul. Using this data we show that (1) volume of inflow of people into a neighborhood is strongly correlated with the diversity of amenities within the neighborhood as measured by the Shannon entropy of the store categories, (2) the volume of inflow of people into a neighborhood is strongly correlated with the productivity of the neighborhood, and (3) the diversity of amenities within a neighborhood is strongly correlated with the diversity of people entering the neighborhood.

Having developed a model that relates diversity of shopping categories, flows of people, and productivity, we then turn to the important problem of predicting the economic growth of neighborhoods. Here we show that (4) the diversity of shopping categories within a neighborhood accurately predicts the economic growth of the neighborhood during the following year. Moreover, the prediction of economic growth using diversity of shopping categories remains accurate even when we control for population density, housing price, and centrality of the neighborhood. Combining (3) and (4) we also find that the diversity of people entering the neighborhood also predicts economic growth, again even when these other variables are controlled.

Finally, we validate this growth prediction model in other cities on two other continents using publicly available data from the social networking and crowd sourcing websites Yelp, Dianping, and Meituan (Chinese group buy and crowd-sourced review sites, respectively). While it is well known that consumer consumption has a positive effect on economic growth[1–3], our work is novel in that it provides a simple quantitative method using openly available data that results in usefully accurate predictions and accounts for a large proportion of variance in growth rates.

Understanding the relationship between shopping diversity, the volume and diversity of flows of people, and economic

productivity and growth can yield insight into a variety of important societal issues and lead to many practical applications such as urban planning. Policy makers and planners who understand the factors that contribute to economic growth in different areas of the city can allocate their resources efficiently in order to make cities grow better and help even out the distribution of wealth across neighborhoods.

While our results do not demonstrate causal processes, they suggest that the diversity of goods and services within a city neighborhood, and the diversity of people flowing into a neighborhood, may be the largest single factor driving human mobility, productivity, and economic growth. These findings suggest that theories of economic growth that emphasize the spread of ideas and opportunities among diverse populations may have stronger causal effects than is generally believed.

# Results

## Flow and Diversity

Experimental results suggest that interactions and information diffusion in societies promote productivity[4–9]. We model how these interactions happen across different regions in a city via a network model of human flow, where the nodes are the regional units (districts, neighborhoods, zip codes, etc.), and the edges represent the volume of flows between the nodes. In order to measure the flow of people across different districts of Istanbul, a set of geo-tagged credit card transactions is utilized. It covers the expenditure of 62,000 people over a period of one year, from June 2014 to June 2015, containing more than four million transactions.

Each transaction is tagged with an exact location, given by its latitude and longitude, and is categorized to one of the 36 different geographically separated districts. We begin by building a directed network of human flows between districts of Istanbul, where nodes are districts. Let the set of individuals that reside in district $i$ be $S_i$ and the set of transactions by person $k$ that occur in distinct places of district $j$ be $T_{kj}$. The network edge weight is then given by:

$$W_{ij} = \sum_{k \in S_i} \left| T_{kj} \right| \tag{1}$$

In other words, an edge is established from district $i$ to $j$ if a resident of district $i$ conducts an in-person transaction in district $j$. The weight of this edge is the total number of such transactions. The empirical probability of flow is then calculated as the fraction of the weights.

Motivated by notion of consumer city by Glaeser et al[10], we investigated how the quantity and diversity of amenities in those regions relate to the different volumes of flows across neighbourhoods. Various places of interest in an urban environment were identified using a dataset created by a commercial digital mapping company. This dataset includes a map of Istanbul with details such as different levels of administrative boundaries (e.g. districts and neighborhoods), and various categories of Points of Interest (POI), published quarterly from 2015 to the end of the first quarter of year 2016. Available POI types in this dataset are provided in the Methods section. The scatter plot in Figure 1 (left) shows that there is a significant relationship between the total volume of inflows and diversity of commodities (measured using Shannon Entropy) in the area, with a correlation value of 0.789. Other than the total volume of inflow, we are also interested in how the quantity and diversity of amenities factor into flows across each pair of districts. As such, we model this inter-district flow of citizens with a gravity model[11] that uses the quantity ($Q$) and diversity ($D$) of amenities as the attractiveness measure and travel times ($d$) between the districts as proximity measure. Under this model, the probability of moving from region $i$ to region $j$ is proportional to the attractiveness of region $j$ and inversely proportional to the proximity of two regions. Formally,

$$P(i \to j) \propto \frac{Q_j^{\alpha} D_j^{\beta}}{d_{ij}^{\gamma}} \tag{2}$$

Using a shared parameter value for the gravity model across all districts, we optimized the value of the parameters $\alpha, \beta, \gamma$, which correspond to the scaling factor of Quantity, Diversity and Distance in the gravity model as shown in equation 2, via simple grid search with a mean squared error objective to minimize. Fig. 1 (right) summarizes the performance of this model. Details of the model can be found in the SI. The mean $R^2$ values of the models is 0.876, with corresponding parameters $\alpha = 0.7$, $\beta = 1.5$ and $\gamma = 1.2$.

The good fit of this model suggests that measurements of attractiveness and not just geographical distance is critical for understanding urban flows. We observe that the optimal model returns an $\alpha$ value of 0.7, which is less than $\beta$, at 1.5, indicating that diversity plays a larger role than the quantity of commodities in determining the attractiveness of a region. Attractiveness scales sub-linearly with quantity, while scaling super-linearly with the diversity. There is also super-linear scaling with the geographical accessibility.

We also fitted independent parameter values for each district, and the district-by-district optimized flow can be found in the SI. Despite the large increase in the degree of freedom of the parameters, the increase in performance, measured by the mean $R^2$ value, is marginal (0.876 to 0.917). This provides evidence for a robust, homogeneous rule that governs the human flow across all districts of Istanbul. This universality across districts is in contradiction with past observations that the rules governing flows vary strongly with physical locations[12].

## Flow and Productivity

Here, we investigate how the volume of human flow can account for the difference in economic productivity in different districts of a city. Existing literature suggests that productivity in urban environments is driven by idea exchange and innovation[6]. Accordingly, we expect that a large number of interactions to arise from large inflows of people, leading to more information flow, innovation, and more economic opportunities for people.

Figure 2 shows the relationships between the flows of each district within a one year period and the corresponding economic productivity (details about the economic indicator are provided in the Method section). The Pearson correlation coefficients between the inflow, outflow, and total flow with the economic indicator are 0.843, 0.475, and 0.840 respectively, indicating a positive relationship between the flows and the districts' economic productivity. We test the significance of the effect of inflow and outflow on economic productivity via linear regressions. The results are shown in Table 1.

The results in Table 1 indicate that the inflow and outflow are significant variables in all models, with a positive relationship with productivity. The results after step-wise inclusion of *residential* population size and within-district flow as independent variables (Models 2, 3, and 4) alleviate the confounding effects of residential population density and flow of the residents within their own district. Moreover, the $R^2$ values of the regression models (at 0.742, 0.743, 0.751 for Models 1, 2, and 3 respectively) show that inflows and outflows explain a large percentage of the variation in the economic productivity of the region. Our results support the hypothesis that the inflow of people and the resulting proximity and interaction with other districts' residents give rise to information diffusion. Additionally, the outflow of residents to other districts allows exploration of new knowledge, which could possibly then be exploited and transferred to the home district upon return. New information then leads to more economic opportunities in the area.

## Consumption Diversity and Flow Diversity

Interestingly, our study shows that there is a significant correlation between the demographic diversity of the inflow of people (see Methods for how it is computed) and the availability of commodities in the region. Figure 3 shows the scatter plot of the diversity of the people (demographics include characteristics such as age, gender, income levels, work, and home districts) entering each district of Istanbul versus the diversity of goods and services available (categorized using the merchant category codes-MCC) in these districts. The two variables have a correlation of 0.825. According to Glaeser et al.[10], availability of diverse goods and services provide the means to attract people with different demographics and varying taste and preferences. Similarly, the inflow of diverse people provides the conditions for a wider variety of businesses to be set up in the region, allowing the local economy to flourish.

## Diversity and Economic Growth

We investigate the above mentioned relationship between the diversity in a neighbourhood, quantified via the Shannon Entropy, and the economic growth. Specifically, diversity is given by the diversity of goods and services consumed in the neighbourhood, as well as the diversity of the inflow of people into the area.

We start by investigating the relationship between diversity of consumption and the growth of the economy. Leveraging publicly available data from social networking and crowd sourcing websites, we are able to extend this part of the study from one city to 3 urban regions in three different countries: Turkey, China, and the United States. These urban regions are divided and studied on a sub-city granularity. Details of the economic indicators and consumption data are defined in the Methods section.

Figures 5A, 6A, and 7A show the scatter plots for regions in Istanbul, Beijing, and various urban areas in the United States. In all three cases, we see that the diversity of consumption exhibits significant statistical positive correlations with the economic growth in the following year at 0.71 (Istanbul, Figure 5A), 0.54 (Beijing, Figure 6A) , and 0.52 (U.S., Figure 7A), indicating that the diversity of consumption alone accounts for between one quarter and half of the variance in future economic growth.

Figures 5B, 6B, and 7B show the residual scatter plots after we account for the correlations with variables such as population density, housing price index, and the geographical centrality of the district within the city. Even after controlling for covariates that could potentially affect economic growth rates, we observe that the diversity still has significant partial correlations with economic growth at 0.72 (Istanbul, Figure 5B), 0.41 (Beijing, Figure 6B), and 0.57 (US, Figure 7B).

We fit an ordinary least squares (OLS) regression model to all available variables. The relationship between economic

growth, $G$ and consumption diversity, $H$ is modelled with the following equation:

$$G_{jt} = \beta_0 + \beta_1 H_{jt} + \beta_2 \rho_{jt} + \beta_3 I_{jt} + \beta_4 D_{jt} + \varepsilon_{jt} \tag{3}$$

where $G_{jt}, H_{jt}, \rho_{jt}, I_{jt}, D_{jt}$ are the growth rate, consumption diversity, population density, housing index, and eigenvalue centrality of the district $j$ at time $t$, respectively.

The regression diagnostics are presented in Table 2. The results show that consumption diversity has a significant, positive effect on growth in all three cases, even after accounting for the effects of covariates such as population density, housing price index, etc. We observe that amongst all the variables, consumption diversity has the most consistent effect across all three regions; it is the only variable that has a consistent direction of dependence and is statistically significant.

We should note that the model is observed to have a poorer fit for the study conducted in the urban areas of the United States, with an $R^2$ of 0.357 (Table 2). This may be attributed to the geographically sparse, and possibly biased, data sample. The Yelp data set we used, merely contains data on a subset of census blocks that were geographically distributed across different states as opposed to the districts studied in China and Turkey, which were all contained within the same city and provided a complete picture of Beijing and Istanbul respectively. In addition, the census blocks are sampled across different time periods $t$. Nevertheless, the regression results in Table 2 show that the positive relationship between consumption diversity and economic growth is reasonably strong and statistically significant.

Additionally, we also find that the inflow diversity has a positive relationship with the economic growth in the neighbourhoods of Istanbul in the following year, as shown in Figure 4. Table 3 shows the results of fitting OLS linear regression models to the data. We find that inflow diversity is a significant predictor of economic growth at a neighbourhood level (Model 1), and this remains true even after control variables are added (Model 2). Interestingly, we find that both inflow diversity and consumption diversity are significant predictors of growth (Model 3) and explain more than half of its variation.

## Discussion

Cities are home to more than half of the world's population and based on the UN projections, cities will attract almost all of the growth in the human population over the next three decades[13]. While cities are known as engines of industry and production, various undesirable factors such as income inequality and socioeconomic segregation affect the citizens' well-being[14–17]. On the other hand, cities, with their high population density, facilitate the consumption of a wide variety of commodities. Thus, understanding how urban environments impact citizens' behaviour in the cities is a research field of increasing importance. This study highlights computational techniques that leverage large-scale datasets to help the planners and policy makers better understand the effect of urban characteristics on economic productivity.

Past work by Glaser et. al[10] has shown that the quantity of amenities is positively correlated with the *population growth rate* of cities, suggesting that people move into cities because they value the diversity of consumption that urban environments provide. Here we extend this idea to the more granular level of individual neighborhoods and districts, and show how commodity diversity predicts both mobility and economic outcomes, by using data on the diversity of amenities in Istanbul and the flow of people within the city.

First we show that human flow between districts of a city can be effectively explained by the relative attractiveness due to the local amenities and ease of access to a neighborhood. A gravity model that takes into account the number and diversity of POIs in the region is capable of predicting more than 85% of the variation in flows between districts. By taking into account the diversity of amenities of the region, we allow the model to take into account intervening opportunities while preserving the effects that physical distance have on discouraging human movement. Additionally, our analysis shows that movement to regions of a city scale super-linearly with the diversity of amenities, in favor of the theory of intervening opportunities[18], suggesting that trip making is not explicitly dependent on physical distance but on the accessibility of resources satisfying the objective of the trip.

Secondly, we show that the differences in the volume of the flow of people into different districts can predict the differences in economic productivity between the various districts of Istanbul. We find that the inflow of people has a strong positive relationship with the economic productivity of the region. Previous works by Bettencourt et al.[14,19] and Gomez-Lievano et al.[20] have shown super-linear scaling of economic productivity with population density at the aggregate level, and Pan et al.[21] argue that the underlying mechanism behind this super-linear scaling phenomena is the establishment of more network ties in denser populations. As the density of the population (both residents and people entering the area for work) increases, the average number of people each city dweller personally interacts with increases as well. The increased rate of information exchange results in more opportunities, which then may lead to increase in productivity observed[4–6,14,22]. Our results fit into this framework[9,23], but in addition quantitatively predicts the variations of economic output in different parts of the city. We find that a linear model that takes into consideration the inflow and outflow of people from a city was able to explain the differences

in economic productivity across different city districts to a high degree (with an $R^2$ value of 0.909). The effects of inflow and outflow of people remained significant predictors of economic output even after accounting for indicators traditionally associated with urban scaling phenomena, such as population. Our results are in favor of this concept that productivity in urban settlements is primarily driven by the strong interaction idea generation exchange within the urban network.

Finally, we show that the consumption diversity and the diversity of human flow, in an urban area can be utilized as a signal about its future economic growth, as there is a positive relationship between the diversity of the neighbourhood and the economic growth in the region a year later, even after controls for population density, housing price index, and the geographical centrality of the district within the city are added.

We propose the following mechanism to explain these results, namely that the flow of diverse people catalyses the production of even more diverse goods and services, which attracts yet more diverse and larger flows of people. Specialised shops and services not only require larger pools of the population but also diverse patrons with different tastes and preferences in order to reach viable levels of demands. These businesses are economically feasible only in diverse, bustling cities. This makes it likely for business owners to invest money in these urban areas, which can support a large variety of specialized businesses, such as niche restaurants that specialize in various cuisines. The result is the increase in the level of economic activity in these areas. Overall, this sets up a dynamic process in which entrepreneurs respond to the flow of people with inputs of money and investments into new businesses. These new amenities provide the conditions, both in terms of their inherent utility and new work opportunities, to continuously attract a flow of people and allow the economy to expand.

One limitation of our study is that the data is not a perfect representation of the flow of people around the cities. Geo-tagged credit card records do not completely capture the movements of people. Only flows that involve economic transactions are captured. Therefore, we would expect the recorded flows to be higher in regions with a higher number of shops, as there is a higher chance of an expenditure occurring based on the sheer numbers alone. A better source of flow data could potentially be from a mobile device equipped with the capability of pinging an individual's location every few minutes. Nevertheless, we mitigated this limitation by including a variable in the model that takes into account the quantity of points-of-interest in the region. Even after accounting for the number of amenities, the diversity of amenities well explained the flow of people.

While we only have direct consumption data in Istanbul, we extended the study to urban areas in two other continents, specifically in Beijing, China and disconnected census blocks in the United States. In these regions, we obtain information about the diversity of consumption via publicly available data. In Beijing, we obtained the data from Dianping, a widely used review platform, and Meituan, which is again a widely used group buy platform. For areas in the United States, we used data available from the Yelp Dataset Challenge. On these platforms, only partial consumption is logged: Meituan only records consumption from shops that offer discount coupons online (though the majority of shops in Beijing offer some sort of deal on the platform). Yelp and Dianping only track consumption that was accompanied by either a review or a check-in. In spite of the use of noisy, open sourced public data from Yelp etc., there is still significant predictive power, suggesting a robust relationship.

Our findings have practical implications for the urban planners and officials of cities. In this new era of Big Data, readily available data sources offer many dynamic measures of a city's economic health from different perspectives[24], and our work demonstrates how electronic records of expenditures can be utilized to do so. City officials need not rely on annualized values of traditional economic indicators for planning purposes but are instead able to obtain up-to-date metrics of how well the city is doing. As a city's ability to remain attractive as a consumption nexus becomes increasingly associated with its success in retaining high-value-adding people and improving its economic well-being, decision makers should consider making their cities a more diverse and vibrant place not just for work, but a pleasant place to live and play in.

## Materials and Methods

### Data

For this study, we utilize various datasets from different sources, including publicly available datasets, data published by governments and census bureaus, datasets from financial sectors, and data provided by a commercial digital map production company.

### Points of Interest Data

The Internet prompted improvements in logistics and supply chain operation, and the omnipresence of online business models means that manufactured goods are national goods that are easily transported. However, cultural goods[25] such as restaurants and theaters are confined to a locality and are representative of the attractiveness of a region as commodities. Therefore, the scope of commodities studied should not be limited to consumer goods. In order to study various types of commodities we use datasets created by a commercial digital mapping company that contains Points of Interest (POIs) found in each neighbourhood. POIs are grouped into sixteen types, namely: business centers, community service centers, financial institutes, educational Institutes, entertainment places, shopping places, restaurants, hospitals, parks, travel destinations, parking lots, auto services, transportation hubs, and level of access to railroads, sectional highways, and major highways.

## Consumption Data

Three data sets are utilized to study urban environments in three different countries: Turkey, China, and the United States.

### *Istanbul, Turkey*

The first data set is a set of geo-tagged credit card transactions in Istanbul that covers the expenditure of 62,000 people over a period of 1 year, from July 2014 to June 2015. It contains more than four million transactions. 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.

### *Beijing, China*

Second is the data on consumer behaviours for people in Beijing, China, which is publicly sourced from the Chinese phone apps Meituan and Dianping. These phone apps are similar to Groupon and Yelp respectively, where users can purchase discount coupons and look up reviews for various amenities in the city. A total of 136,000 deals offered by 6,500 food businesses for four months, (November - December 2016, and April - May 2017) are considered. A total of two million customers are in this data set.

### *United States*

Third is the publicly available Yelp Data Challenge dataset that we use to study the consumption patterns of individuals in the United States. This dataset includes records on check-ins, reviews, and ratings from 654,135 users of 26,149 businesses across 42 different discontinuous urban areas in the United States from 2009 to 2015.

## Theoretical Framework

## A Predictive Model of Human Flows

### *Flow Network*

Understanding the dynamics of economic growth in cities has been a very important research field in the context of urban studies. Experimental results suggest that interactions as a result of human flow to different regions in cities promote productivity. Examples of these theories include the theory of structural holes[4], weak ties[5], the effect of social interaction on economy flourish[6,7], and the importance of information flow in the Research and Development[8,9]. Motivated by these research studies[4–9,26–28], we propose a network model of human flow and leverage various datasets for this purpose.

Here we use a notion of directed network to show the flows across different regions of a city. In this directed network, region units are the nodes and an edge is established from region $i$ to $j$ if a resident of region $i$ visits a distinct point of interest in region $j$. The points of interest can be of different type such as workplace, retail store, or restaurant. Visits from other regions to region $i$ are called in-degree centrality or in-flow ($IF_i$), and visits made by residents of region $i$ to other regions are called out-degree centrality or out-flow ($OF_i$). For the Istanbul case, we consider each district of Istanbul as a node and obtain empirical values of in-flow and out-flow using the credit card transaction dataset. The network edge weight $W_{ij}$ is defined as the total volume of flow from $i$ to $j$, given by the equation in 1. This approach results in a fully connected directed network with 36 nodes, and 1296 edges.

### *Gravity Model*

Our aim is to understand and predict the flow of citizens between different districts using credit card and map data, and investigate the factors associated with the volume of flows. Based on the definition of flow given above, people visit different places in a region either in relation to their work or for other purposes such as shopping and visits with hedonistic motivations[29]. We contend that the number of amenities as well as their diversity could be a region's attractiveness for citizens. In statistical physics, Shannon entropy of a system is a measure of the number of permutations of the states of the system[30]. In the context of urban amenities, it measures the number of different ways to combine these amenities together. Therefore it represents a metric of the variety of experiences that city dwellers obtain from this mix of amenities. Moreover, ease of access to the region of interest is another important factor to attract more visitors[31]. Thus, we use the Huff gravity model[11] to model the flows between districts. Huff model is a technique in spatial analysis which is based on the principle that the probability that a consumer in district $i$ will visit and purchase something in district $j$ is a function of the relative distance $d_{ij}$ and attractiveness of district $j$. In this model, the flows are governed as follows. For an individual that originates from district $i$, each district $j$ has a utility $U_{ij}$ that is proportional to its attractiveness $A_j$ and inversely proportional to the distance $d_{ij}$ from the origin.

$$U_{ij} = \frac{A_j}{d_{ij}^{\gamma}} \tag{4}$$

In particular, we used the product of POI count, $Q$ and POI diversity, $D$ (raised to appropriate powers) as the measure of attractiveness for each district. The diversity of POIs in district $j$ is measured by the Shannon Entropy of the distribution of POI types $P_k^{(j)}$.

$$A_j = (Q_j)^{\alpha}(D_j)^{\beta} \tag{5}$$

$$D_j = \sum_k -P_k^{(j)} \log\left(P_k^{(j)}\right) \tag{6}$$

Instead of Euclidean distance, we use the travel time via public transportation as the distance $d_{ij}$. This is a more realistic measure of the ease of access. We use Google Maps API to get these travel times. Google Map Distance Matrix API allows users to obtain the average travel time between two distinct points by entering the origin and destination coordinates. We use the credit card transaction dataset to in turn determine the central origin and destination points in each district. Origin for each district is calculated as the average of residents' home locations and destination is the average location of merchants in those districts.

Under this model, the proportion of trips originating from $i$ that will have $j$ as its destination, $\hat{f}_{ij}$, is proportional to its utility relative to other competing districts.

$$\hat{f}_{ij} = \frac{U_{ij}}{\sum_{j'} U_{ij'}} \tag{7}$$

To the best of our knowledge, this study is the first to model the shopping area choice with the Huff model and the proposed novel area attractiveness measure with number and diversity of urban amenities.

### *Optimizing parameters*

Each gravity model that describes the flows of district $i$ is parameterized by $\alpha$ , $\beta$ , $\gamma$ , and we optimize these parameters using a 3-d grid search, selecting the parameters that correspond to the smallest mean squared error $(MSE)$.

$$MSE(\alpha,\beta,\gamma) = \frac{1}{|J|}\sum_{j\in J}(f_{ij} - \hat{f}_{ij})^2 \tag{8}$$

## Measuring Economic Productivity

We attempt to measure the economic productivity of the city of Istanbul via the Gross Domestic Product (GDP). Since official GDP records from Turkey are only published with granularity at a city level, we use the insurance sales data as a proxy for GDP at a finer district level. This dataset is provided by a Turkish Insurance Company that has been ranked one of the top three in sales volume based on different insurance categories. The dataset contains the company's total contract sales values in all categories in Istanbul districts level, during 2014, 2015, and 2016.

The proxy used is the sum of insurance sales in each district. Among all the insurance categories, we consider four categories, namely: Home, Workplace, Vehicle liability, and Vehicle Collision insurances. The reason why these insurance types were chosen is that all of them are based on physical assets like buildings or vehicles, which is a more direct measure of economic output. In order to test whether the chosen insurance types can be used as proxy for GDP, the district-level insurance sales in the chosen four categories for 80 cities in Turkey for the year 2014 was summed up and compared with published GDP of those 80 cities in the same year. The evaluation metric used was Pearson's correlation, which was found to be 0.99249 between two sets (more details are provided in supplementary information). The result shows that insurance valuations published could serve as a proxy for GDP at the district level.

## Inferring Consumption Diversity from Data

We begin by obtaining the set of local businesses that exist within each regional unit $j$, $S^{(j)}$ . These sets of local businesses are divided based on their categories $k$ into subsets $S_k^{(j)}$. These categories include types of restaurant, such as fast-food, Japanese, Mexican, etc., or types of facilities like parks, movie theatres, etc., based on the nature of the businesses. Additional information about the regional units and business categories are provided in the supplementary information.

From credit card expenditures in Istanbul and the Meituan dataset in China, the consumption of a good or service is directly obtainable. For the Yelp and Dianping data sets the rate of consumption of a good or service $i$, $C_i$ is estimated by adding up the

total number of customers who have either checked in, rated, left a review, or posted a picture about the amenity. Thus, for each region $j$, we are able to obtain the proportion of consumption in a particular category $k$

$$P_k^{(j)} = \frac{\sum_{i \in S_k^{(j)}} C_i}{\sum_k \sum_{i \in S_k^{(j)}} C_i} \tag{9}$$

We also define a metric of diversity of urban amenities consumed by people in region $j$ using the Shannon entropy[32]

$$D(POI_j) = \sum_k -P_k^{(j)} \log\left(P_k^{(j)}\right) \tag{10}$$

By obtaining different sets of amenities that were consumed by users across different time periods, we are able to investigate the relationship between attractiveness of local goods and economic growth across different years.

## Diversity of the Inflow of People.

For each individual $k$, we have a vector $X_k$ that describes the demographics of an individual. For the study of Istanbul, $X_k \in \mathbb{R}^6$, where the vector $X_k$ contains information on the **home district, work district, age category, gender, education level, and income deciles** of the individual.

We compute $q_x^{(j)}$, the proportion of in-flows in district $j$ associated with people with a particular demographic $x$ . The demographic diversity of people entering district $j$ is thus given by

$$D_j = \sum_{x \in X} -q_x^{(j)} \log\left(q_x^{(j)}\right) \tag{11}$$

## Measuring Economic Growth

### *Istanbul*

We measure growth of Istanbul via changes in its GDP between 2014 and 2016.

$$G_{jt} \approx \frac{GDP_{j,t+1} - GDP_{jt}}{GDP_{jt}} \tag{12}$$

### *Beijing*

We measure the economic development in region $j$ of Beijing at time $t$, $C_{jt}$, 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 of future economic growth.

$$G_{jt} \approx \frac{C_{j,t+1} - C_{jt}}{C_{jt}} \tag{13}$$

### *United States*

Similar to Turkey, we measured the growth of counties via changes in its economic output. Again, since official GDP records are only published at a state level, we approximated it via the sum of personal incomes in each county $j$, which is obtained via the United States census bureau.

$$G_{jt} \approx \frac{\text{Total Personal Income}_{j,t+1} - \text{Total Personal Income}_{jt}}{\text{Total Personal Income}_{jt}} \tag{14}$$

## Growth Model

We propose the following model in order to measures the relationship between economic growth rate ($G$), and various factors in urban area:

$$G_{jt} = \beta_0 + \beta_1 H_{jt} + \beta_2 \rho_{jt} + \beta_3 I_{jt} + \beta_4 D_{jt} + \varepsilon_{jt} \tag{15}$$

where $H_{jt}, \rho_{jt}, I_{jt}, D_{jt}$ are consumption diversity, population density, housing index, and eigenvalue centrality of the district $j$ at time $t$ respectively. The control variables are explained below.

### *Population Density*

The population density $\rho_{jt}$ is defined as the number of people per unit area (in $1000/km^2$).

### *Housing Index*

Rental prices act as a medium that regulates investment into the region. Increase in rental prices may discourage businesses to be physically set up in the region, or reduce the supply of labor as individuals are reluctant to move into the area. To account for the effects of property prices on growth, $I_{jt}$ is included as a covariate, which is defined as the ratio of the current rental price (per unit area) to the cheapest rental in the first time period $t_0$.

$$I_{jt} = \frac{\text{Rent}_{jt}}{\text{Rent}_{j^* t_0}} \tag{16}$$

where $j^* = \arg\min_j \text{Rent}_{jt_0}$.

### *Eigenvalue Centrality in a Geographical Network*

The success of a district is also dependent on its location in the city. Generally speaking, the more central the location, the easier the access to the businesses, and therefore, the more likely it would attract customers. To control for the location factor, we first computed a geographical network of districts with the edges between districts $i$ and $j$ weighted by the reciprocal of the travel distance (in measured in minutes) between them. We then computed the eigenvalue centrality of each business in this geographical network[33].

## Data and Code Availablity.

The data and code used in the analysis can be found on https://github.com/cshikai/Cities.

# References


1. Glaeser, E. L. & Gottlieb, J. D. Urban resurgence and the consumer city, DOI: 10.1080/00420980600775683 (2006).
2. Clark, T. N., Lloyd, R., Wong, K. K. & Jain, P. Amenities drive urban growth. *J. Urban Aff.* DOI: 10.1111/1467-9906.00134 (2002).
3. Clark, T. N. 3. Urban Amenities: Lakes, Opera, and Juice Bars: Do they drive development? 103–140, DOI: 10.1016/S1479-3520(03)09003-2.
4. Burt, R. S. Structural holes: The social structure of competition. *Harv. Univ. Press. Camb. Massachussetts* DOI: 10.1177/0265407512465997 (1992).
5. Granovetter, M. S. The Strength of Weak Ties. *Am. J. Sociol.* DOI: 10.1086/225469 (1973).
6. Granovetter, M. The Impact of Social Structure on Economic Outcomes. *J. Econ. Perspectives* DOI: 10.1257/0895330053147958 (2005).
7. Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E. & Pentland, A. Mining face-to-face interaction networks using sociometric badges: predicting productivityin an IT confiriguration task. In *IS Winter Conference*, 1–19, DOI: 10.2139/ssrn.1130251 (2008).
8. Allen, T. J. *Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information Within the R&D Organization* (1984).
9. Reagans, R. & Zuckerman, E. W. Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams. *Organ. Sci.* DOI: 10.1287/orsc.12.4.502.10637 (2001).
10. Glaeser, E. L. & Maré, D. C. Cities and skills. *J. Labor Econ.* **19**, 316–342, DOI: 10.1086/319563 (2001).
11. Huff, D. L. A Probabilistic Analysis of Shopping Center Trade Areas. *Land Econ.* **39**, 81, DOI: 10.2307/3144521 (1963).
12. Isaacman, S. *et al.* A tale of two cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, HotMobile '10, 19–24, DOI: 10.1145/1734583.1734589 (ACM, New York, NY, USA, 2010).
13. Crane, P. & Kinzig, A. Nature in the metropolis. *Science* **308**, 1225, DOI: 10.1126/science.1114165 (2005).
14. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kuhnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci.* **104**, 7301–7306, DOI: 10.1073/pnas.0610172104 (2007).
15. Clark, W. A. Residential segregation in American cities: a review and interpretation. *Popul. Res. Policy Rev.* **5**, 95–127, DOI: 10.1007/BF00137176 (1986).

16. Trounstine, J. *Segregation by Design: Local Politics and Inequality in American Cities* (2018).

17. Chen, Y. & Rosenthal, S. S. Local amenities and life-cycle migration: Do people move for jobs or fun? *J. Urban Econ.* DOI: 10.1016/j.jue.2008.05.005 (2008).

18. Stouffer, S. A. *et al. Intervening opportunities: a theory relating mobility and distance* (1940).

19. Bettencourt, L. & West, G. A unified theory of urban living, DOI: 10.1038/467912a (2010).

20. Gomez-Lievano, A., Patterson-Lomba, O. & Hausmann, R. Explaining the prevalence, scaling and variance of urban phenomena. *Nat. Hum. Behav.* **1**, DOI: 10.1038/s41562-016-0012 (2017).

21. Pan, W., Ghoshal, G., Krumme, C., Cebrian, M. & Pentland, A. Urban characteristics attributable to density-driven tie formation. *Nat. Commun.* **4**, DOI: 10.1038/ncomms2961 (2013).

22. Pan, W., Aharony, N. & Pentland, A. Fortune monitor or fortune teller: Understanding the connection between interaction patterns and financial status. In *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing*, 200–207, DOI: 10.1109/PASSAT/SocialCom.2011.163 (2011).

23. Quigley, J. M. Urban Diversity and Economic Growth. *J. Econ. Perspectives* DOI: 10.1257/jep.12.2.127 (1998).

24. Glaeser, E. L., Kominers, S. D., Luca, M. & Naik, N. Big data and big cities: The promises and limitations of improved measures of urban life. *Econ. Inq.* **56**, 114–137, DOI: 10.1111/ecin.12364.

25. Kloosterman, R. C. Cultural amenities: Large and small, mainstream and niche - a conceptual framework for cultural planning in an age of austerity. *Eur. Plan. Stud.* **22**, 2510–2525, DOI: 10.1080/09654313.2013.790594 (2014). https://doi.org/10.1080/09654313.2013.790594.

26. Dong, X. *et al.* Social Bridges in Urban Purchase Behavior. *ACM Trans. Intell. Syst. Technol.* **9**, 33:1—-33:29, DOI: 10.1145/3149409 (2017).

27. Centola, D. & Macy, M. Complex Contagions and the Weakness of Long Ties. *Am. J. Sociol.* DOI: 10.1086/521848 (2007).

28. Zhou, X., Hristova, D., Noulas, A., Mascolo, C. & Sklar, M. Cultural investment and urban socio-economic development: A geosocial network approach. *Royal Soc. Open Sci.* DOI: 10.1098/rsos.170413 (2017). 1806.03378.

29. Arnold, M. J. & Reynolds, K. E. Hedonic shopping motivations. *J. Retail.* DOI: 10.1016/S0022-4359(03)00007-1 (2003).

30. Jaynes, E. T. & Mechanics, S. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630, DOI: 10.1103/PhysRev.106.620 (1957).

31. Sim, A., Yaliraki, S. N., Barahona, M. & Stumpf, M. P. H. Great cities look small. *J. The Royal Soc. Interface* **12**, 20150315, DOI: 10.1098/rsif.2015.0315 (2015).

32. Shannon, C. E. A mathematical theory of communication. *The Bell Syst. Tech. J.* **27**, 379–423, DOI: 10.1145/584091.584093 (1948).

33. Newman, M. E. J. The mathematics of networks. *The New Palgrave Encycl. Econ.* **2**, 1–12, DOI: 10.1057/9780230226203.1064 (2007). 1408.5240.


For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file.

## Acknowledgements


We thank the MIT Trust Data Consortium and Prof. Guan at the Capital Institute of Science and Technology Development Strategy, Beijing for making this project possible. We also thank X. Dong , Y. Leng and Y. Yuan for helping with the collection and cleaning of the Meituan/Dianping data set and the various companies, who wish to remain anonymous, for providing the insurance data, credit card data in Istanbul, and housing prices in Istanbul.


## Author contributions statement

SKC collected data, performed research and analyzed data for consumption and economic growth; MB collected data, performed research and analyzed data for commodities and human flow; HC collected the economic data for Beijing; SB collected the data for Istanbul; and SKC, MB, AP and BB contributed to writing.

## Additional information

**Competing interests**. The authors declare that they have no competing financial interests.

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Inflow | 21.080*** | 21.712*** | 20.553*** | 20.512*** |
| | (2.6715) | (3.7222) | (2.7105) | (3.8990) |
| Outflow | 14.752*** | 16.831* | 20.064** | 19.955* |
| | (5.1697) | (9.8939) | (7.1530) | (10.344) |
| Within-district Flow | | -4.132 | | 0.253 |
| | | (16.675) | | (17.202) |
| Population | | | -0.386 | -0.397 |
| | | | (0.36984) | (0.387) |
| Constant | -153476.8 | -164618.5 | -92124.5 | -91219.5 |
| | (107900) | (118342) | (121931) | (138224) |
| Observations | 36 | 36 | 36 | 36 |
| $R^2$ | 0.7426 | 0.7433 | 0.7517 | 0.7517 |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 1.** Regression coefficients for prediction of Economic Productivity using Human Flows

| | Beijing | Istanbul | United States |
|---|---|---|---|
| Consumption Diversity, H | 0.233*** | 0.707*** | 0.583*** |
| | (0.0493) | (0.121) | (0.0706) |
| Population Density, $\rho$ | 0.474*** | -0.389* | -0.213** |
| | (0.0644) | (0.464) | (0.0729) |
| Housing Index , I | 0.164*** | -0.113 | 0.0981 |
| | (0.0469) | (0.147) | (0.0703) |
| Geographic Centrality, D | 0.222*** | 0.245 | 0.270*** |
| | (0.0635) | (0.209) | (0.0727) |
| Constant | -1.97e-09 | 4.11e-09 | -2.19e-09 |
| | (0.0425) | (0.117) | (0.0678) |
| Observations | 187 | 36 | 145 |
| $R^2$ | 0.671 | 0.574 | 0.357 |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 2.** Regression coefficients for prediction of Economic Growth using Consumption Diversity

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Inflow Diversity, H | 0.5929*** | 0.7714*** | 0.4378* |
| | (0.141) | (0.132) | (0.196) |
| Consumption Diversity, H | | | 0.3955* |
| | | | (0.180) |
| Population Density, $\rho$ | | -0.6869*** | -0.5497** |
| | | (0.464) | (0.188) |
| Housing Index , I | | -0.1685 | -0.1629 |
| | | (0.148) | (0.140) |
| Geographic Centrality, D | | 0.3066 | 0.271 |
| | | (0.208) | (0.197) |
| Constant | -2.97e-05 | 4.11e-09 | -2.19e-09 |
| | (0.139) | (0.117) | (0.110) |
| Observations | 36 | 36 | 36 |
| $R^2$ | 0.342 | 0.576 | 0.635 |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 3.** Regression coefficients for prediction of Economic Growth using Inflow Diversity

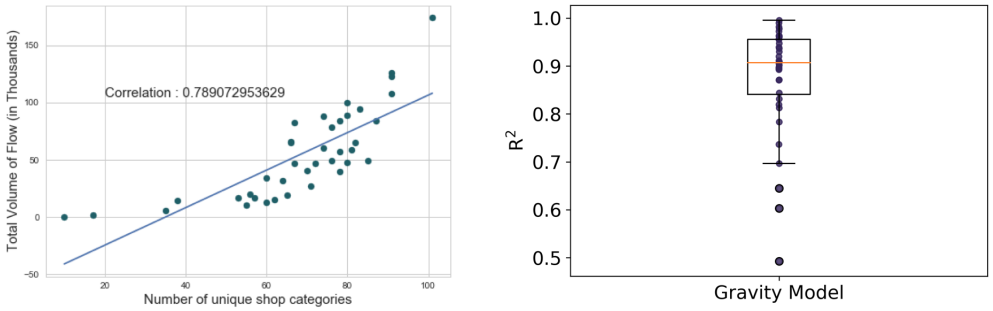


**Figure 1.** Left - Scatter plot of the number of unique shop categories versus total inflow volumes of each district in Istanbul. Right - Distribution of the $R^2$ values of predicted flows, overlaid with box plot.

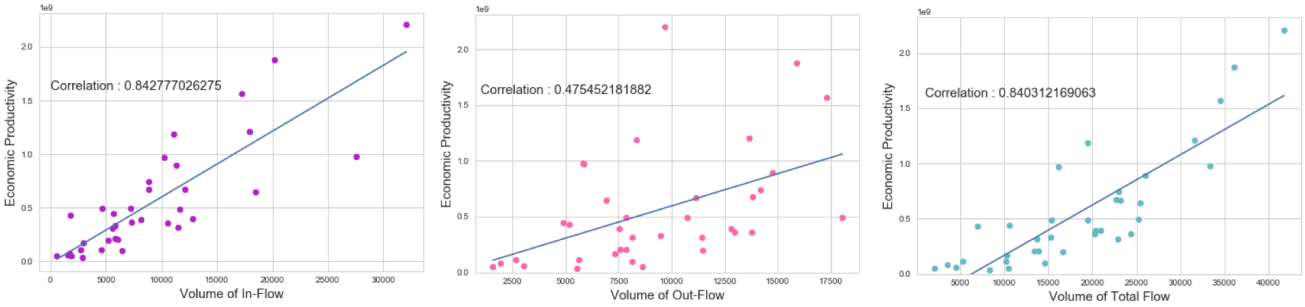


**Figure 2.** Scatter plot of total flow volumes (Left : Inflow, Middle : Outflow, Right : Total Flow) and the economic productivity of each district in Istanbul.

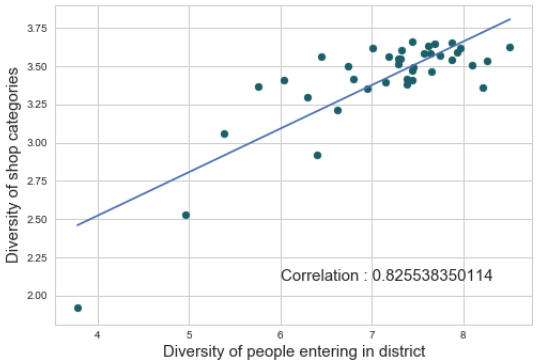


**Figure 3.** Scatter plot between the demographic diversity of people entering each district and the diversity of amenities available in each district of Istanbul.

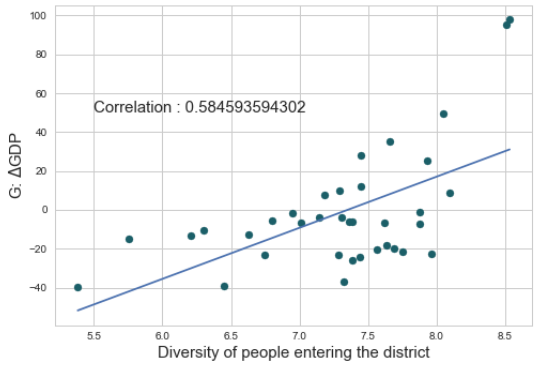


**Figure 4.** Scatter plot between the demographic diversity of people entering each district and economic growth.

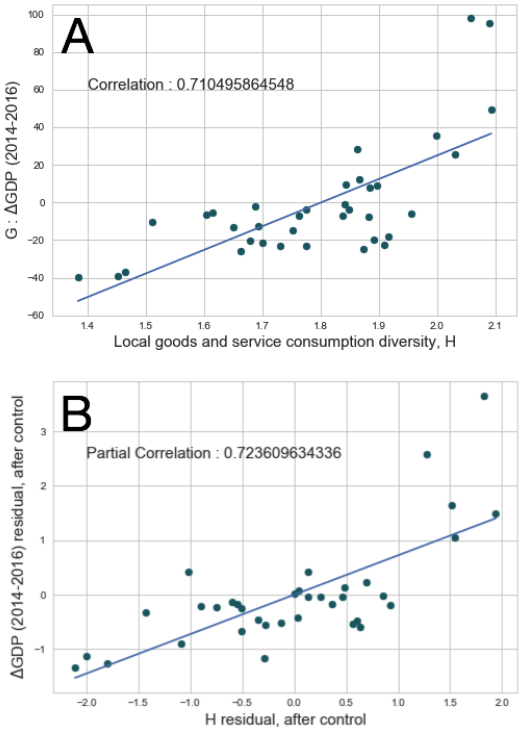


**Figure 5.** Scatter plot of growth vs diversity of goods consumed. Each data point corresponds to a district in Istanbul in a specific year. Top (A) - Full correlation. Bottom (B) - Partial correlation after controlling for other variables.

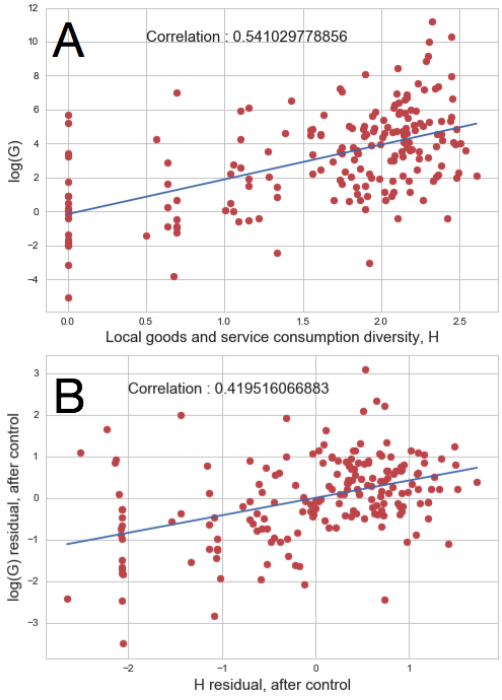


**Figure 6.** Scatter plot of growth vs diversity of goods consumed. Each data point corresponds to a district in Beijing in a specific year. Top (A) - Full correlation. Bottom (B) - Partial correlation after controlling for other variables.

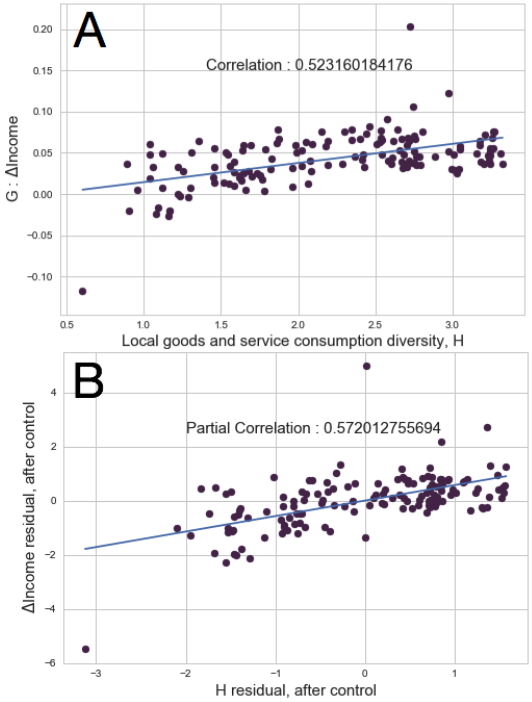


**Figure 7.** Scatter plot of growth vs diversity of goods consumed. Each data point corresponds to a census block in United States in a specific year. Top (A) - Full correlation. Bottom (B) - Partial correlation after controlling for other variables.