RESEARCH

Analyzing and Predicting the Spatial Penetration of Airbnb in U.S. Cities

Giovanni Quattrone[1*], Andrew Greatorex[2], Daniele Quercia[3], Licia Capra[2] and Mirco Musolesi[2]

[*]Correspondence:
g.quattrone@mdx.ac.uk
[1]Middlesex University, London, UK
Full list of author information is available at the end of the article

Abstract

In the hospitality industry, the room and apartment sharing platform of Airbnb has been accused of unfair competition. Detractors have pointed out the chronic lack of proper legislation. Unfortunately, there is little quantitative evidence about Airbnb's spatial penetration upon which to base such a legislation. In this study, we analyze Airbnb's spatial distribution in eight U.S. urban areas, in relation to both geographic, socio-demographic, and economic information. We find that, despite being very different in terms of population composition, size, and wealth, all eight cities exhibit the same pattern: that is, areas of high Airbnb presence are those occupied by the "talented and creative" classes, and those that are close to city centers. This result is consistent so much so that the accuracy of predicting Airbnb's spatial penetration is as high as 0.725.

Keywords: quantitative analysis; spatial data mining; sharing economy; Airbnb

Introduction

Airbnb is a hospitality service that allows people to rent their unused rooms or entire properties, by directly engaging in computer-mediated transactions with potential guests. Instead of being based on centralized entities, this example of peer-to-peer (a.k.a. sharing / collaborative / asset) economic model is based on a distributed network of individuals directly accessing each other underused assets (in this case, accommodations). Founded in 2008, Airbnb has grown exponentially in the past few years [1, 2, 3], and now it has over 3,000,000 listings in more than 65,000 cities across the globe [4].

Critics say that the rapid growth of Airbnb has been accelerated by a lack of regulation. This has given rise to political and regulatory debates about *how* to best compile legislation for businesses utilizing Airbnb's model of collaborative consumption. In the field of Law, researchers have indeed made the case for regulating Airbnb. Stephen Miller, for example, has put forward the idea of legalising short-term rental markets like those enabled by Airbnb via "transferable sharing rights" [5], with each house owner being given the right to engage in short-term rental for a given period of time a year. But how should these rights be allocated, and how should they be priced? Since the actual dynamics behind Airbnb penetration have so far received little attention, there is not much evidence upon which to build policies.

To support evidence-based policy making, we study the relationship between Airbnb's penetration in a variety of cities and each city's geographic, demographic, and socio-economic characteristics. In so doing, we set out to answer two main questions: *(i)* Which factors explain Airbnb spatial penetration in urban areas?; and *(ii)*

Can we predict Airbnb spatial penetration in a city using historical data from other ones? In answering those questions, we make three main contributions:

- We gather data about Airbnb listings for eight U.S. cities. These are Austin, Los Angeles, Manhattan (New York City), New Orleans, Oakland, San Diego, San Francisco, and Seattle. We have chosen them because they vary in size, population composition, wealth, and cost of living.
- We propose a method for explanatory analysis of geographic data, and study the relationship between Airbnb's spatial penetration and geographic, social and economic conditions in these eight cities.
- We find that, despite being very different, most of the cities considered in this study show the same pattern: high level of penetration is associated with central locations and with presence of talented and bohemian people, which some scholars refer to as the 'creative class' [6, 7]. These relationships are statistically strong, so much so that we are able to build a predictive model for Airbnb's spatial penetration that generalizes across cities and that has an accuracy as high as 0.725.

These results suggest that a generic geographic penetration prediction model for Airbnb might be applied across different cities. Such model can be particularly helpful to policy-makers. Indeed, new phenomena like Airbnb do not penetrate all cities at the same time – i.e., some cities will act as early adopters, while many others will follow later. If adoption in a later-coming city could be predicted using a generic geographic model built from observations of early-adopting cities, then municipalities could pro-actively deploy policies to direct adoption and growth in selected areas based on models' estimates. For example, if we refer back to the "transferable sharing rights" scheme by Miller, one could consider allocating more rights and at lower cost to house owners in areas located further from the city center, since these areas tend to naturally suffer from low Airbnb penetration; viceversa, owners in central areas and with high concentration of people working in the creative industries could be given fewer rights and at higher cost, since our analyses revealed that hot-spots of Airbnb rentals are invariably linked to such areas. We will expand on this subject later in the paper, when we discuss potential policy implications of our study.

Related Work

Our work relates to the growing literature on the sharing economy, which has been carried out in a variety of disciplines, from Law to Economics, from Sociology to Computer Science. Overall, previous work has focused on two main themes: the impact of Airbnb on the hospitality industry, and whether and how Airbnb should be regulated.

Studies on Sharing Economy Platforms

Researchers have recently started to study the social dynamics behind service platforms. They studied, for example, the role of geographic factors (e.g., geographic distance and population density) in the success of two service platforms: Uber (ride-hailing app) and TaskRabbit (an app for hiring people to do things such as assembling flat pack furniture) [8]. They also looked into how socio-economic factors were

associated with the use of Uber in Seattle [9] and found that a neighborhood's racial composition mattered [10].

As for Airbnb, scholars have begun to examine its relationship with more traditional forms of hospitality, and that has added yet more nuance to the critical debate that continues to surround the platform. Zervas *et al.* [11] analyzed Airbnb listings in Texas, and found that Airbnb had negatively affected the revenue of local hotels: a 1% increase in the number of listings led to a 0.05% loss of hotel revenue. However, that mainly impacted lower-end hotels and left untouched higher-end ones. In another study, Varma *et al.* [12] found hotels and Airbnb listings to be quite complementary (e.g., customers tend to be different) and, as such, they concluded that Airbnb hardly creates disruption in the industry. More recently, Quattrone *et al.* [13] looked at the growth of Airbnb in London over four years and found that growth started in central areas as expected by touristic demand, but then moved to socio-economically deprived areas as well – people were likely renting their spare rooms to make ends meet. Our work builds upon that previous research and makes two main new contributions. It tests for two previously overlooked aspects: robustness and generalizability. Unlike previous research (focused on London), our work reports on *robust* findings: among all considered metrics, we identify a subset of them showing a consistent relationship with Airbnb penetration across the eight U.S. cities analyzed. Our work also goes beyond a descriptive analysis by offering a predictive validity that is *generalizable*: the proposed predictive model uses historical data of Airbnb's penetration in $(N-1)$ cities to estimate Airbnb's penetration in the N^{th} city, and has an accuracy as high as as 0.725.

Proposals for Regulation

Edelman and Geradin proposed a few ways in which platforms such as Airbnb could be regulated without compromising market efficiency for both consumers and service providers [14]. In a similar vein, Koopman *et al.* [15] made a case for policy change. Einav *et al.* [16] took a systematic approach and spelled out pros and cons of a few cities' current regulations. Stephen Miller proposed to legalize the short-term rental market through "transferable sharing rights" [5], where each house owner has the right to engage in a short-term rental for a given period of time. This right can be sold to others, if the owner does not wish to engage. The revenues generated by the sharing right market would go to both the city council, which would be able to raise revenues without raising taxes any further; and to neighborhood groups, which would be compensated for any externality. Then, to ensure market efficiency, web platforms could sell transferable sharing rights in a way similar to what StubHub does when selling tickets. Other academics have taken a more hands-off stance, however. Being an innovation, Airbnb should not be excessively regulated, Ranchordas argued [17]. The general principle behind his proposal is that innovations should not be stifled by regulation. Cohen and Sundararajan opted for self-regulatory approaches and for leaving regulatory responsibility to parties other than the government [18]. Finally, to regulate the sharing economy, one should understand what 'sharing' means in that context. Zale offered a taxonomy of 'sharing', including formality and gratuity [19], and Ikkala and Lampinen showed that Airbnb transactions are not purely financial – they are mediated by hospitality and sociability [20].

Those proposals have remained within academic circles, and critics say that Airbnb's rapid growth has been nurtured by a severe lack of regulation. One often cited concern is that of revenues from tourism. Tourists have long been a source of income for governments (through taxes) and hotels alike. As the sharing economy (e.g., Airbnb) continues to expand in unregulated areas, not only traditional industries but also governments are bound to suffer [21]. On the other hand, proponents have put forward evidence of the contrary: they claim that peer-to-peer markets have been found to create wealth, stimulate optimal resource utilization, and even reduce environmental impact [22].

To move the debate forward, we need a systematic study that looks at Airbnb presence in relation to geographic, social and economic conditions of urban areas, to establish what conditions are associated with Airbnb penetration (and the lack thereof). With such evidence at hand, legislators can then develop policies to help control Airbnb growth in areas of certain characteristics.

Research Questions

In this paper, we analyze the spatial penetration of Airbnb in cities across the U.S., with the aim of answering two main research questions:

- *RQ1. Which factors explain Airbnb spatial penetration in urban areas?* We investigate a comprehensive range of geographic, social, and economic characteristics of urban areas, and quantify their relative importance in capturing the spatial penetration of Airbnb offerings in such areas. We are particularly interested in investigating whether there exists a small set of common characteristics that are linked to Airbnb presence, across a variety of cities.
- *RQ2. Can we predict spatial penetration in a city from the dynamics observed in other cities?* We investigate the possibility of developing a generalized prediction model based on the characteristics identified above, so that we can accurately predict the Airbnb penetration in a U.S. city, having trained the model on other American cities. If so, legislators in a city where Airbnb is only starting to appear can use our model to forecast areas of (potentially undesirable) under/over Airbnb penetration. This will enable them to put in place policies early on, aimed at steering its growth where desired.

Next we illustrate the datasets and the metrics we have used to answer these research questions.

Datasets and Metrics

Cities

We analyzed eight different cities located within the U.S.. These vary substantially among each other – for example, in terms of size, population composition, and wealth, as described later in this section. We chose to focus on the U.S., as this country hosts a diverse range of cities, with mature Airbnb presence across many of them. Future studies may wish to explore to what extent the findings that hold within a country also span across different ones.

San Francisco. It is the city where Airbnb was founded in 2008 and is currently headquartered. As Airbnb's hometown, it offers insights into the most developed Airbnb marketplace. Furthermore, it is the second most densely populated U.S. city

and is home to many technology entrepreneurs who work in the nearby heart of the U.S. technology scene, the Silicon Valley. It is a very ethnically diverse city, has a very high average age and, despite having high median income, has a large disparity between the rich and poor.

Oakland. Unlike San Francisco, it serves as a center for trade and is the busiest port in California. Despite its close proximity to San Francisco, the characteristics of Oakland's demographic makeup differ considerably and median pay is roughly two thirds that of San Francisco's.

Manhattan. It is the most densely populated borough of New York City. It is also the city's economic and administrative center, and it is often described as the cultural and financial capital of the world. Manhattan has the highest cost of living in the U.S., and also contains the country's most profound level of income inequality. The majority of the population is white (65%), and approximately 27% are foreign born.

New Orleans. In stark contrast to Manhattan, New Orleans is the smallest of the chosen cities, with a population of 378,000, predominantly black (60.2%). The city has seen a decline in population in recent times. As further proof of contrast to Manhattan, the median income of the city is $26,900 (2010 U.S. Census), to Manhattan's $72,200, almost three times greater.

Austin. It differs vastly to both the metropolis of Manhattan and the quaint New Orleans. Austin is the fastest growing city of the top 50 largest U.S. cities and is not so ethnically diverse. The majority of Austin's population is white (66.8%). It is also the youngest city in the dataset.

Seattle. The Pacific Northwest city of Seattle, in Washington, is an important center for technology, being home to Amazon, Microsoft, and Boeing. It is also a major gateway for trade with Asia. Like Austin, it is a predominantly white city. However, it is far older, has a much higher median income, and a greater cost of living.

San Diego. It is the third major city in our dataset (with a population greater than 1,000,000). The city, which has an immediate proximity to the Mexican border, is not a technology hub like Seattle or New York. Its main economic engines are the military and tourism. Due to its closeness to Mexico, it has a large Hispanic population and a low proportion of black people (6.7%).

Los Angeles. It is a global center of commerce and has a diverse economy in business, technology, culture and sport. It has the highest educational diversity in the country and ranks highly on the diversification of its economy business-sectors. Despite its size and economic power, it has a low median income and a disproportionately high cost of living.

Table 1 lists the eight cities chosen for this study (first column), and also summarizes their varying social and economic characteristics (next five columns), in terms of: population, median age, median income, percentage of white population, and cost of living – estimated from consumer prices of goods and services relative to the reference urban area of Manhattan [23].

Airbnb Data

We gathered the Airbnb data made available on Murray Cox's website as part of his "Inside Airbnb" project (`http://insideairbnb.com/`). The website periodi-

Table 1: Summary characteristics of the 8 chosen U.S. cities.

City	*Pop*	*Median Age*	*Median Income*	*% of White*	*Cost of Living*	*# Airbnb Listings*
Austin, TX	885k	31	$32k	67%	74	5,193
Los Angeles, CA	3.8m	34	$28k	50%	81	17,044
Manhattan, NY	1.6m	36	$72k	65%	100	16,041
New Orleans, LS	379k	35	$27k	33%	80	2,646
Oakland, CA	406k	35	$32k	35%	88	1,155
San Diego, CA	1.4m	36	$33k	60%	78	3,530
San Francisco, CA	837k	39	$49k	49%	99	6,361
Seattle, WA	652k	36	$43k	68%	91	2,711

Population, Median Age, Median Income, Percentage of White are derived from the official U.S. Census Bureau published in 2010. Cost of Living has been derived from `https://www.numbeo.com/cost-of-living/region_rankings.jsp?title=2017&region=019`. Finally, the number of Airbnb Listings has been determined from our Airbnb dataset.

cally publishes snapshots of Airbnb listings around the world. On May 2016, we downloaded all the *listings* in our eight cities. We also verified the validity of the data by selecting 10 random listings in each city and double checking both their presence on the original Airbnb platform, and the accuracy of their locations. Location accuracy is key for us as we measure *Airbnb penetration* as the number of Airbnb listings in a given area. The last column of Table 1 reports the number of Airbnb listings for each of the eight selected cities.

Explanatory Variables

In order to explain the varying spatial penetration of Airbnb within U.S. cities, we looked into three different groups of variables, capturing their *geographic*, *social* and *economic* context. Most of these variables have been obtained from the most recent U.S. Census bureau (that is, Census 2010 – `https://www.census.gov/`) which gathers decennial population data. The U.S. Census data is available at a census *tract* spatial granularity; that is, the smallest territorial unit of analysis for which population data is available in the U.S.. Other sources of data include: OpenStreetMap (`https://www.openstreetmap.org`), Google Maps, and a variety of official city websites (as summarized in Table 2).

City Geography

Distance to Center. A previous Airbnb study [13] of the city of London, UK, found that distance to the city center was one of the variables that most explained Airbnb presence in an area (i.e., the closer to the city center, the more Airbnb listings). We aim to explore whether the same holds for U.S. cities. Some of the analyzed cities (such as San Diego, Oakland and Seattle) are relatively small with a clear definition of city center. For other cities this may be not true and they may contain multiple urban hubs [24]. For simplicity, we computed a single metric across all cities; specifically, we consider the 'downtown district' or CBD (central business district) as the center of the city. For each city, we compute distance to center as the shortest distance in meters between the CBD's center, and the center of the tract under study.

Points of Interest. A point of interest (POI) is a geographic feature that might be useful or interesting. Examples of POIs include pubs, town halls and post offices.

A study of the geography of Airbnb in London [13] found that, together with 'distance to center', the 'tourism factor' of an area, as shown by the number of POIs within an area, had the greatest positive significance on the number of Airbnb offerings in that area. We expect that the relationship will hold for American cities too, such that areas of higher POI concentration, indicating greater tourist appeal, will also have increased Airbnb presence. To count the number of POIs within a given area, we used OpenStreetMap data; specifically, for each city, we extracted the latitude/longitude coordinates for all POIs that fell under the following OpenStreetMap categories: accommodation, attractions, eating and drinking, retail and sports, and entertainment.

Number of Hotels. Despite a previous analysis showing that in London there is little relationship between hotels and Airbnb adoption [13], we do not know *a priori* whether the same conclusion holds in U.S. cities as well. Airbnb's economic blog, which reports and measures Airbnb's effect on city economies, states that 72% of Airbnb properties in San Francisco are outside the central hotel district [25]. However, little other evidence exists relating the spatial penetration of Airbnb listings to that of hotels. Intuitively, the number of hotels in an area should provide a reasonable proxy for the level of tourism of that area. Furthermore, results highlighting where Airbnb listings appear in a city relative to hotels will provide regulators with a source of quantitative information to make more informed decisions. We thus explore this variable in our analysis. Since there is no publicly available dataset for the number of hotels in all cities, hotel data was crawled from Google, searching for 'city_name' + 'hotels', and then retrieving their latitude-longitude pairs.

Bus Stops. The strength of an area's infrastructure and transport links have historically been a key component in the performance of property prices, due to the ease of connection to major areas of that city. For tourists visiting a city, although they may spend time and money in tourist centers, their choice of where they stay is likely influenced by the connectivity of an area. Different cities may offer a variety of different public transport modalities. Since buses are present in all cities under study, we chose the number of bus stops in an area as proxy to the strength of said area's transport links. Thus, we expect to see a relationship between Airbnb offerings and the number of bus stops. To compute this metric, we used a combination of OpenStreetMap data and city-specific datasets to obtain the latitude-longitude of bus stops; we then counted the number of stops within each area.

Population Density. This is a standard metric derived from the U.S. Census Bureau that provides information on how densely populated a specific area is. It is widely used as general statistical datum at the country as well as at the local level. It is calculated by dividing the number of people living in a certain area by the area's total surface. Population density is an aspect considered crucial by many urbanists in explaining a number of urban aspects [26, 27, 28]. Recent studies have found that this factor is linked to the spread of sharing economy services [29]. We thus decided to include it as one of our geographic attributes.

Social Indexes

Race Diversity Index. The Race Diversity Index is a metric derived from the U.S. Census Bureau; it provides a measure of how much racial diversity exists in an

area. First coined by Meyer and Macintosh [30], it is formulated as a Gini-Simpson Index [31] and acts as a probability measure. It measures the likelihood that two people selected at random from a given area represent different types. In this case, it is a measure of whether the race of the chosen people is the same. We formulate the problem with seven distinct racial categories: white, black or African American, Hispanic or Latino, American Indian or Alaska native, Asian, native Hawaiian or Pacific Islander. The greater the race diversity index, the greater the probability that two people selected at random will be from different races.

Income Diversity Index. The income diversity index shows how diverse an area is in terms of average household income for the population of that area. It is derived from the U.S. Census Bureau and it is calculated using the Gini-Simpson index [31] for three distinct wage bands: low income (annual incomes less than $35,000), middle band income (annual incomes between $35,000 and $100,000) and high income (annual incomes greater than $100,000).

Bohemian Index. A bohemian is a socially unconventional person with interests in art or literacy (`https://en.oxforddictionaries.com/definition/bohemian`). Richard Florida's paper "Bohemia and Economic Geography" [6] examines the relationship between geographic concentrations of bohemia and a strong technology presence by directly measuring the bohemian population at an MSA (Metropolitan Statistical Area) level. Though there are other variations of the bohemian index [32], we use Florida's definition, which computes the proportion of the number of bohemians to the number of residents in an area, compared to the national proportion of bohemians to the number of the total population. We derived the Bohemian Index from the U.S. Census Bureau.

Talent Index. The talent index [33] measures the education level of a populace, defined as the proportion of people with a bachelor's degree or above. The index is normalized per thousand people and it is derived from the U.S. Census Bureau. Richard Florida hypothesizes that a high talent index is correlated with a larger concentration of bohemians. Given this, we may infer that areas with a strong technology presence, such as those areas with high Airbnb uptake, will have a higher index for talent.

Proportion of Young People. This was calculated as the proportion of people aged between 20 and 34 years old in a given area against the population of that area. Florida suggests that, as well as the bohemian index, a higher concentration of young people is often a driver of the technology uptake in that area [6]. We derived this index from the U.S. Census Bureau.

Economic Indexes

Unemployment Proportion. The unemployment proportion is calculated as the number of people aged 16 and over currently out of work (unemployed) against the total number of people in an area. This measure is provided by the U.S. Census Bureau. Unemployment rates often provide a strong indication of the economic health of an area. According to Florida's work on the 'creative class' [7], areas of lower unemployment (amongst other factors) are symbolic of a creative class, and transitively may lead to greater technology concentration. We would thus expect to see a negative correlation between Airbnb penetration and unemployment proportion. However,

the Wall Street Journal [34] found a large percentage of Airbnb renters were offering up living spaces due to unemployment. In Paris, only one third of Airbnb hosts were reported to have full time jobs [35]. If the relationship holds across the U.S. too, then we may see a positive correlation between unemployment and Airbnb penetration instead.

Poverty by Income Percentage. Michael Zweig [36] defines poverty as "a state of deprivation, or a lack of the usual or socially acceptable amount of money or material possessions". In the U.S., the most common poverty metrics are the 'poverty thresholds', as defined by the U.S. Census Bureau [37]. Our explanatory variable is then calculated, in a given tract area, as the percentage of households in poverty (as defined by their income) against the total number of households in that area. The underlying hypothesis is that Airbnb's penetration will fall in areas of increased poverty.

Median Household Income. For each tract area, the U.S. Census Bureau measures the median household income for the local population. A temporal study on Airbnb in London [13] showed that income became increasingly more negative correlated with Airbnb penetration over time, signaling that more people with low income were joining Airbnb as hosts, possibly using the extra income generated from Airbnb to support themselves.

Median Household Value. The U.S. Census Bureau also provides a measure of median household value for each area. Together with median household income, this variable should provide a strong indicator of socio-economic makeup of a city. This can also be used to identify clusters of cities with similar profiles.

Proportion of Owner Occupied Residences. Quattrone *et al.* [13] found that, in London, Airbnb hosts tend to rent rather than own the property. Therefore, we hypothesize that the proportion of owner occupied residences matter in the U.S. as well. We derived this metric from the U.S. Census Bureau.

Table 2 summarizes all the metrics introduced in this section, along with the sources from which they were taken.

Method

In this section, we first define the spatial unit of analysis that was adopted throughout the study. We then outline the methods used to answer each of our research questions.

Spatial Unit of Analysis

To quantify the relative importance of geographic, social and economic factors within a city with respect to Airbnb listings, we first need to define a spatial unit of analysis. We chose to operate at the level of *tracts*, the smallest granularity at which the U.S. Census Bureau collates data, for three main reasons: first, since each tract has roughly the same population (about 4,000 inhabitants) [38], the adopted metrics – such as number of hotels and number of POIs – do not need further normalization (i.e., they are implicitly normalized by the number of people residing in that area). Second, census tracts cover a contiguous area; if this were not the case, it would be difficult to measure spatial autocorrelation by analyzing clusters and dispersion of data. Third, census tracts represent a unit of measurement that captures a statistically significant number of data points. All metrics summarized in Table 2 have been computed at tract level.

Table 2: Dataset summary.

Category	*Acronym*	*Metric*	*Source*	*Description*
Airbnb	bnb_penetration	Airbnb penetration	insideairbnb.com	Number of Airbnb listings registered in a given city area
Geography	distance	Distance to Center	Routing Machine	Distance from the center of a given area to the center of a city
	poi	Points of Interest	OpenStreetMap	Number of points of interest (POI) in an area
	hotel	Number of Hotels	Google Maps	Number of hotels in an area
	bus	Bus Stops	OpenStreetMap, Official city websites	Number of bust stops in an area
	popDens	Population Density	Census	Population density in a given area
Social	race_div	Race Diversity Index	Census	Diversity of races in an area
	income_div	Income Diversity Index	Census	Diversity of income in an area
	bohemian	Bohemian Index	Census	Proportion of people employed in arts, entertainment and media in an area to the same proportion nationwide
	talent	Talent Index	Census	Proportion of people in an area with degrees higher than an associate degree
	young	Proportion of Young People	Census	Proportion of people aged between 20 and 34 to the total populous of the area
Economic	unemployment	Unemployment Ratio	Census	Proportion of unemployed to the total populous
	poverty	Poverty by Income Percentage	Census	Proportion of the number of residents with income in poverty to the number of residents with income in an area
	income	Median Household Income	Census	Median estimate of household income in an area
	household_value	Median Household Value	Census	Median value of a household in an area
	owner	Proportion of Owner Occupied Residences	Census	Proportion of dwellings that are owned to those that are occupied

RQ1. Explanatory Analysis

Our first research question investigates whether it is possible to explain Airbnb spatial penetration using geographic, social and economic variables. To find what variables are significantly correlated with Airbnb penetration and to what extent, we use a multivariate linear regression model in the form of Ordinary Least Squares (OLS)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon \tag{1}$$

where y denotes the Airbnb penetration in a given area; $x_1, \cdots, x_k$ are the set of the explanatory variables that reflect the geographic, social, and economic conditions of the same area (see Table 2); $\beta_0, \beta_1, \cdots, \beta_k$ are the unknown parameters; and ϵ is the error term.

Since some of our metrics are skewed and therefore do not conform with the normality assumption of the variance, we compensate for the skewness of such variables by applying a log transformation. Further, since our metrics are on very different scales, we standardize them by computing their z-scores. This transformation enables us to compare β coefficients that are from different distribution scales.

An issue to consider at this stage in the chosen method is that of 'multicollinearity' [39, 40], that is, the possibility for two or more explanatory variables to be correlated with each other. In the presence of multicollinearity, the variance of the standardized β coefficients increases. As a result, although the regression models

are still valid overall, their detailed interpretation is much more difficult (e.g., the more correlated the explanatory variables are, the more difficult it is to determine how much variation in the outcome each separate explanatory variable is responsible for). To tackle this issue, we used a technique called Variance Inflation Factors (VIF) [39, 40]. The VIF associated with an explanatory variable is obtained by, first, performing the linear regression with the explanatory variable as dependent variable and the other remaining variables as independent ones and, second, by using the overall model fit (i.e., the R^2 value) obtained in the previous step in the following formula

$$VIF = \frac{1}{1 - R^2}. \tag{2}$$

If the explanatory variable has a strong linear relation with at least another one, the corresponding model fit is likely to be close to 1, and the explanatory variable's VIF to be large. Various recommendations for acceptable levels of VIF have been offered in the literature; as an example, a value of 10 is commonly recommended as the maximum level of acceptable VIF [41, 39, 42, 40]. Candidate variables showing a VIF higher than the recommended maximum level must be excluded from the list of predictors of the regression model.

Finally, for each of our OLS regression models, we need to test for their validity. In general, regression models assume that explanatory variables are independent of one another. However, since we are mostly dealing with geographic data, this assumption might be violated. This is because geographic data often obeys to Tobler's First Law of Geography: "Everything is related to everything else. But near things are more related than distant things" [43]. This means that the spatial features we use in our regression models (e.g., number of POIs, household income) may tend to be geographically clustered together. If this is the case, we run the risk of under-estimating the chances of committing Type I errors, and being drawn to the conclusion that our explanatory metrics and model are better at explaining variation in Airbnb penetration than they actually are. To test the validity of our regression models, we use the well-known Moran's I [44] to measure spatial autocorrelation of the residuals in our OLS models. To calculate Moran's I, we needed a clear definition of "nearby" observations. We chose one of the most common ways of computing spatial connectivity matrices [45]: we first computed the latitude and longitude of the central point of each census tract; then, we computed the spatial connectivity matrix as the inverse Euclidean distance of these points.

RQ2. Predictive Analysis

Our second research question investigates whether it is possible to predict the spatial penetration of Airbnb listings in a city, based on what has been learned about their spatial penetration in other cities. In other words, we investigate the generalizability of the findings concerning RQ1. To this purpose, we use classification analysis.

To begin with, based on the previous regression analysis, we identify a subset of geographic, social and economic variables that are most important across the eight cities under exam. By most important, we mean they are statistically significant across most cities (i.e., p-values lower than 0.05), and that exhibit a consistently high

β score (in absolute value). In other words, we identify a minimum set of variables that our predictor model will use. In so doing, we also reduce the dimensionality of our dataset, thus reducing the chance of over-fitting. We then scale the selected variables, so that each of them describes how many standard deviations it exceeds (or it is below) its mean value for a given city. This step is necessary since the cities in our dataset have different characteristics and, as such, it is not possible to compare them in absolute terms; for example, the median household income in Manhattan is not comparable at all to that of New Orleans.

For our target variable, that is, the number of Airbnb offerings per tract, we stratify it into categorical values, to form a discrete set of labels: 'low', 'medium' and 'high' penetration. Choosing the right technique that transforms continuous data in bins is a non-trivial process. Since our target variable of Airbnb penetration shows a very skewed distribution (i.e., many U.S. tracts have very low Airbnb penetration, and only few of them have high penetration), we ended up using a logarithmic binning [46] that produced the following bins: the 'low' penetration bin containing around 70% of lowest values of the target variable; the 'medium' penetration bin containing around the next 20% values of the target variable; and the 'high' penetration bin containing roughly the remaining 10% of observations. Figures 1 and 2 show the result of this binning strategy; specifically, Figure 1 shows the frequency distribution of Airbnb penetration in the eight U.S. cities analyzed against the three obtained bins; Figure 2 shows the Choropleth Maps of Airbnb penetration according to the same three bins.

On the transformed data, we compute eight different instances of our model, one for each city under study. We use the data for the city under consideration as test set and the data related to the remaining seven cities as training set. In terms of the classification algorithms used, we experimented with: *(i)* Support Vector Machines (SVM) with radial basis function kernel, a classifier generally used when the number of features is greater than the number of training examples, as it is in our case; *(ii)* logistic regression, a classifier that works on the assumption of data linearity and that is not negatively affected by data collinearity; *(iii)* random forest, a classifier that generally works without any assumption of data linearity and collinearity; and *(iv)* Naive Bayes, a classifier that, although assumes conditional data independence, it has been found to perform well in practice. We used the following metric to assess the quality of our classifications

$$weighted\ accuracy = \frac{c \times TP + TN}{c \times (TP + FN) + (TN + FP)}. \tag{3}$$

This accuracy metric is very suitable when the classes to be predicted are imbalanced [47, 48], as it is in our case. In this formula, c is equal to the class imbalance, that is, the extent to which the negative class is more frequent than the positive one; TP, FP, FN and TN indicate, respectively, the true positive, false positive, false negative and true negative classification cases. With such a definition of accuracy, a trivial "always predict the most common class" classifier would achieve a weighted accuracy equal to 0.5. Therefore, any classifier having a weighted accuracy higher that 0.5 improves over both a random guess fare and a trivial classifier with imbalanced classes.

Table 3: Analysis of Airbnb offering.

		Austin	*Los Angeles*	*Manhattan*	*New Orleans*
Geography	distance	-0.63	-0.26	-0.24	-0.30
	poi	-0.04	0.04	0.22	0.14
	hotel	-0.02	0.04	-0.04	0.03
	bus	0.13	-0.01	0.08	-0.06
	popDens	-0.10	0.00	-0.13	0.05
Social	race_div	-0.22	0.02	-0.13	0.16
	income_div	0.01	0.08	0.05	0.08
	bohemian	0.16	0.27	0.51	0.25
	talent	0.00	0.31	0.04	0.35
	young	0.00	0.07	0.28	0.04
Economic	unemployment	0.04	0.05	0.00	-0.02
	poverty	0.11	0.04	0.14	0.19
	income	-0.06	0.10	-0.35	-0.05
	household_value	0.09	0.27	0.01	0.01
	owner	-0.03	-0.19	-0.08	-0.07
	Adjusted R-squared	0.70	0.75	0.60	0.70
	Moran's I	0.06	0.05	0.03	0.04

		Oakland	*San Diego*	*San Francisco*	*Seattle*
Geography	distance	-0.06	-0.35	-0.07	-0.09
	poi	-0.05	0.07	0.19	0.30
	hotel	-0.10	0.04	-0.03	0.04
	bus	0.20	0.06	0.29	0.01
	popDens	-0.09	0.10	0.22	0.01
Social	race_div	0.09	-0.15	-0.10	0.28
	income_div	0.07	-0.03	0.11	0.04
	bohemian	0.10	0.07	0.24	0.11
	talent	0.70	0.47	0.37	0.22
	young	0.14	0.21	0.13	0.20
Economic	unemployment	0.01	0.06	-0.03	-0.09
	poverty	0.07	0.10	0.20	-0.12
	income	-0.28	0.03	0.18	0.08
	household_value	0.40	0.14	-0.06	0.21
	owner	-0.09	0.00	0.10	-0.15
	Adjusted R-squared	0.74	0.70	0.62	0.47
	Moran's I	0.04	0.03	0.06	0.02

We fit the models for the 8 different cities separately, one for each city. Each cell indicates the standardized β coefficient of the model for the corresponding city. Blue bars are associated with positive β coefficients, red bars to negative ones. The shade of the bars encode p-values: dark bars are associated with p-values lower than 0.05, light bars are associated with p-values greater than or equal to 0.05.

We compare the performance of our classifications against a benchmark model, which is based on the single most significant variable identified in RQ1. We define the variable of most significance to be the largest absolute standardized regression coefficient of the prior regression analysis, averaged over all eight cities.

Results

RQ1. Explanatory Analysis

We begin by presenting the results of the regression models. Table 3 shows β coefficients for each variable associated with Airbnb penetration for each of the eight analyzed cities, along with the adjusted R^2 and Moran's I values for each model. Furthermore, the β values are accompanied by blue and red bars, representing the size and sign of the coefficient; blue bars represent positive coefficients and red bars represent negative ones. The most important variables in each model are those with the highest absolute beta values (longest bars). Finally, p-values are chromatically visualized using the colors of the shade of the bars: dark blue/red bars indicate p-value values less than 0.05; conversely, light blue/red bars indicate p-value values greater than or equal to 0.05.

First and foremost, the results may not be significant if they present evident spatial autocorrelation. We find that all models are robust under this aspect; i.e., we did not find evidence that results are based on spatial auto-correlative factors (Moran's I ≤ 0.06).

Geography. As far as geographic factors are concerned, the results show that distance from the city center has a strong, negative relationship with the Airbnb offering across 5 over 8 cities. That is, the further away a tract is from the city center, the fewer the number of Airbnb establishments. Only in Oakland, San Francisco and Seattle the distance from the center is not considered as one of the most important variables. Additionally, the attractiveness of an area, characterized by the number of points of interest (POIs) is statistically positively correlated with the number of listings for 5 out of 8 cities. This is indicative that Airbnb listings are predominantly located in more touristic areas. Finally, the number of bus stops per tract was positively correlated with Airbnb penetration in the three cities of Austin, Oakland, and San Francisco.

Social Indexes. The bohemian and talent indexes exhibit strong positive correlation with Airbnb offerings across the selected U.S. cities, with the former being as high as 0.51 in Manhattan, and the latter being 0.70 in Oakland. This finding is in agreement with Florida's research [6, 7] and is further substantiated by the βs of the number of young people, which similarly follows a cross-city pattern of positive correlation. Income diversity does not bear a significant relationship with Airbnb offerings instead. Race diversity has strong correlation, but of opposite sign across different cities: it is negatively correlated with Airbnb offerings in Austin and San Diego, but positively correlated in New Orleans and Seattle. This suggests that different dynamics are taking place between Airbnb offerings and race diversity in different U.S cities, and a universal model cannot capture them.

Economic Indexes. In terms of economic indexes, we find that, despite not playing a predominant role in each model, the median household value is positively correlated with Airbnb penetration in five out of eight U.S. cities analyzed. At the same time, we find that the median income of an area is inversely correlated in Manhattan and Oakland (although not significant in most other cities). Taken together, a possible explanation for this is that Airbnb adopters are renting out rooms in houses they do not own.

The above results suggest that, in U.S. cities, Airbnb listings are predominantly clustered in tracts that are close to city centers and touristic attractions, and that are home to a young, creative and talented crowd. Since the regression models across cities show a good fit (adjusted R^2 are consistent and high – they range between 0.47 and 0.75, with an average value of 0.66), we hypothesize we can take the most significant variables and use them to build a generalizable predictive model. However, before we present our predictive analysis results, there are important concerns still relating to our explanatory analysis that we need to discuss first, starting with the issue of multicollinearity.

Multicollinearity

So far, in studying the standardized β coefficients, we did not consider 'multicollinearity', that is, that the explanatory variables might be correlated with each

other. Yet, as Figure 3 shows, some variables are indeed highly correlated with each other, with the highest conditional dependencies appearing for variables 'talent', 'bohemian', 'income' and 'household_value'.

Multicollinearity in the regression model might inflate the β coefficients and compromise their interpretability casting doubts on the interpretations we previously offered of our explanatory variables. To deal with this issue, we applied the VIF technique [39, 40] described in the Method section, and initially accepted all variables with a maximum VIF level not greater than 10. All explanatory variables satisfied this condition, suggesting our previous results were correctly interpreted. Some scholars though are more restrictive and suggest lowering the maximum accepted VIF to 4 [49, 50]. We tested this more restrictive threshold, and found that only one explanatory variable among 'poverty', 'talent', 'income', 'bohemian', and 'household_value' could be kept without suffering from multicollinearity. We then built five different sets of regression models, each set containing only one of the flagged variables (i.e., set_1 contains 'poverty', set_2 contains 'talent', set_3 contains 'income', set_4 contains 'bohemian', and set_5 contains 'household_value'). Each of these sets contains 8 model instances, one for each city under study. We then examined the results of the regressions, to verify whether the same explanatory variables that were statistically significant in the full model (Table 3) were also confirmed significant (and with the same sign) in these restricted models.

Table 4 shows the aggregated results. Specifically, each column corresponds to one set of models; each row corresponds to one explanatory variable within the models, with the first 10 variables (from 'distance' to 'owner') being present in all model sets, since they did not have VIF higher than 4, and the last five (from 'poverty' to 'household_value') being present in one model set only. Each cell in the table then indicates the number of standardized β coefficients that are statistically significant and higher than 0.15 in absolute terms; such number is positive if there are more instances (i.e., cities) within that model set with the variable having a (significantly) positive coefficient, and negative otherwise. As an example, the number '–7' in the first row of column set_1 indicates that, for the corresponding set of models, the standardized β associated with Distance to Center is statistically significant and lower than -0.15 in 7 cities out of 8.

Table 4: Aggregated results for five different sets of models.

	set_1	set_2	set_3	set_4	set_5
distance	–7	–5	–6	–6	–6
poi	+5	+4	+5	+5	+5
hotel	–1	0	0	–1	–1
bus	+2	+3	+4	+2	+3
popDens	–1	+1	+1	+1	0
race_div	–3	–2	–1	–2	–1
income_div	+3	+1	+3	+1	+4
young	+4	+3	+5	+4	+6
unemployment	–1	0	0	–2	0
owner	–1	–2	–3	–2	–1
poverty	–2	na	na	na	na
talent	na	+6	na	na	na
income	na	na	+4	na	na
bohemian	na	na	na	+8	na
household_value	na	na	na	na	+6

In each set of models only one variable among 'poverty', 'talent', 'income', 'bohemian', and 'household_value' is included in the computation.

Let us consider first the variables that we previously found significant and that did not have multicollinearity issues (i.e., they are present in all model sets): by looking at Table 4 and Table 3, we confirm that distance from the center is (significantly) negatively correlated with Airbnb penetration in most cities across all model sets, while presence of a young population and presence of POIs (tourist attractions) are positively correlated. If we then look at the variables that were previously found significant but that were flagged for multicollinearity, we now find that in the model containing Bohemian Index (set_4), such variable is confirmed to be significantly positively related with Airbnb penetration across all the eight U.S. cities analyzed; likewise, in the model containing Talent (set_2), such variable is confirmed to be significantly positively related with Airbnb penetration in six out of eight U.S. cities analyzed. Such consistency of results strengthens the validity of the results presented for the overall model.

Sensitivity Analysis

Our Airbnb data dates back to May 2016, while the official U.S. Census data – which is the latest one – dates back to 2010. We argue that, despite the misalignment of six years, the two sets of data can be analyzed in combination, since census conditions do not change significantly in six years. Indeed, if one correlates each variable in the 2000 U.S. census data with the same variable in the 2010 census (census data is updated every 10 years), then the resulting Spearman correlations are quite high (Figure 4) with only a few exceptions (namely, Diversity Index, Bohemian Index, and Unemployment Ratio).

To measure how sensitive our results are, relative to census data change, we re-computed our regression models, now extracting explanatory variables from U.S. Census 2000, and comparing results against when extracting variables from U.S. Census 2010. Table 5 shows the results obtained: for each row (i.e., for each of the 15 explanatory variables of the full model), we count the number of cities in which the variable was found significantly (positively/negatively) correlated with Airbnb presence using U.S. Census 2000 (first column), and when using U.S. Census 2010 (second column). Results are strikingly similar, suggesting that our model is robust against (past) changes in census data (although we cannot speculate what would happen for future census data changes). One explanatory variable for which the results appear to change is 'owner'. That is because areas with uptake in ownership tend to benefit from increases in Airbnb adoption (we found a Spearman correlation equal to 0.35 between $\text{owner}_{2010} - \text{owner}_{2000}$ and bnb_penetration – p-value< 0.001). Areas with high Airbnb presence today are areas where there was significantly less ownership 16 years ago; fast forward a decade, residents have increasingly bought properties in such areas, so they now engage with Airbnb rentals.

Comparison with Airbnb Penetration in London, U.K.

Beside this study, the only work that to date investigates the relationship between Airbnb's spatial penetration and geographic, social and economic conditions in a city is the one proposed by Quattrone *et al.* [13] where the investigation was conducted in London. It so appears to be an interesting opportunity to relate our findings to those obtained in London. Below we report all the commonalities and differences between

Table 5: Comparison of results obtained by using explanatory variables derived from U.S. census 2000 against results obtained by using U.S. census 2010.

		Census 2000	*Census 2010*
Geography	distance	−6	−5
	poi	+3	+3
	hotel	0	0
	bus	+2	+2
	popDens	+1	+1
Social	race_div	+1	0
	income_div	+2	0
	bohemian	+6	+5
	talent	+3	+5
	young	+1	+2
Economic	unemployment	−1	0
	poverty	−1	+1
	income	−1	−2
	household_value	+2	+2
	owner	−3	0

our results and those illustrated in [13] across geographic, social and economic factors (summarized in Table 6).

Geography. Our results strongly match those illustrated in [13] for London in three different key aspects: *(i) distance to center*, *(ii) tourism factor* and *(iii) hotel presence*. Specifically, our results show that Distance to Centre has a strong, negative relationship with the Airbnb penetration across 5 out of 8 cities; unsurprisingly, exactly the same finding is discovered also in London. Furthermore, our results show that the 'tourism factor' of an area – measured as the density of certain types of points of interest related tourist attractions – is positively correlated with Airbnb penetration in 4 out of 8 cities. Findings illustrated in [13] back-up this hypothesis; in this last case, the 'tourism factor' was measured as the density of Foursquare check-ins considered as a rough proxy of how many tourists each area attracts. Finally, both our findings and those reported in London confirm that there is no relationship between hotel presence and Airbnb adoption. Despite these big commonalities there are also some different trends between the geographic factors correlated with the Airbnb adoption in the eight analyzed U.S. cities and in London. Specifically, our findings show that in 2 out of 8 U.S. cities analyzed the presence of infrastructure and transport in an area is also an indicator of higher Airbnb penetration. This finding is not supported in London, possibly because the public transport offering is more homogeneous across the capital.

Social. Social factors are those exhibiting the strongest differences between London and the 8 U.S. cities analyzed instead. Specifically the three factors most positively correlated with Airbnb penetration in the eight U.S. cities analyzed are *bohemian* and *talent* and, in a few cities, the presence of *young* people. In London, only the presence of young people is positively correlated with Airbnb penetration; surprisingly, the bohemian index does not correlate with Airbnb penetration in London, whereas the talent index was not considered in the study. We speculate that this discordant trend is due to the different demographic makeups of American cities – where racial segregation and demographic divides are often high[1] – as opposed to those of London.

Economic. We have found a strong agreement between our findings and those illustrated in [13] in London in: *(i) unemployment*, both findings confirm that there

[1] https://fivethirtyeight.com/features/the-most-diverse-cities-are-often-the-most-segregated/

is no relationship between unemployment and Airbnb adoption; *(ii) income*, which is negatively correlated with Airbnb adoption both in London and in two of the eight analyzed U.S. cities; *(iii) household_value* , which is positively correlated with Airbnb adoption both in London and in two of the eight analyzed U.S. cities. Despite these big commonalities there are also some slight different trends between the economic factors correlated with the Airbnb adoption in the eight analyzed U.S. cities and in London. Specifically, in London there is a statically significant negative relation between the proportion of owner occupied residences and Airbnb penetration. Our analysis only partially confirms these results; in fact, the Proportion of Owner Occupied Residences is negatively correlated with Airbnb penetration in two out of eight analyzed cities; however, these correlations do not appear to be statistically significant (see Table 3).

Table 6: Comparison of results between the eight analyzed U.S. cities and London, UK.

		Eight U.S. cities	*London*
Geography	distance	–5	–
	poi	+3	+
	hotel	0	0
	bus	+2	0
Social	race_div	0	0
	bohemian	+5	0
	young	+2	+
Economic	unemployment	0	0
	income	–2	–
	household_value	+2	+
	owner	0	–

Each cell under the column 'Eight U.S. cities' indicates the number of standardized β coefficients that are statistically significant and higher than 0.15 in absolute terms; such number is positive if there are more instances (i.e., cities) within that model set with the variable having a (significantly) positive coefficient, and negative otherwise. Each cell under the column 'London' indicates whether the corresponding correlation was significant positive (+), or significant negative (–), or rather not significantly correlated (0). The variables 'popDens', 'income_div', 'talent' and 'poverty' are not reported since the same parameters were not considered in the study proposed by Quattrone *et al.* [13].

RQ2. Predictive Analysis

We built a predictive model by first selecting the variables with the highest statistically significant β coefficients: distance from the center, POI, bohemian, talent, income, household value, young and population density. For the benchmark model, we used as predictor only the variable with the highest statistically significant β – that is, distance from the center. We then followed the method we previously proposed to answer RQ2. Figure 5(a) shows the weighted accuracies obtained by our classifiers in the eight cities, averaged for the three classes of 'Low', 'Medium' and 'High' Airbnb penetration; Figure 5(b) shows the same weighted accuracies obtained by our benchmark.

In comparison to the chosen benchmark, our full model outperforms it for all classifiers and all cities. However, the accuracy of the prediction strongly depends on the chosen classification method, it varies from city to city and not all the penetration rates are equally easy to be estimated.

Full models Vs. benchmark. Our model yields a weighted accuracy ranging between 0.58 (San Francisco – logistic classifier) to 0.72 (San Diego – random forest classifier). Conversely, the benchmark yields an accuracy ranging between 0.49 and 0.60. The best results for the benchmark are obtained for New Orleans and Austin, where the benchmark is close to the full model, implying that distance to center is an extremely important factor in these cities.

Results for classification methods. The best overall accuracies are obtained by the random forest classifier for both our model (weighted accuracies ranging from 0.61 to 0.72) and the benchmark (weighted accuracies ranging from 0.50 to 0.60), suggesting that it is beneficial to account for non-linearity and interaction effects.

Results for the different cities. Among all cities, we achieve the best accuracy for Seattle and San Diego – the best weighted accuracy was obtained for San Diego using the random forest classifier (weighed accuracy equal to 0.72). The cities that were most difficult to estimate are San Francisco and Manhattan, where the random forest classifier reaches a weighted value equal to 0.61 for San Francisco. This is perhaps to be expected, given that Manhattan and San Francisco are very diverse. We conjecture that, due to high population concentrations,their tracts encompass a multitude of diverse socio-economic characteristics that cancel out, to a certain extent, expected patterns.

Results for penetration rates. Figure 6 shows the weighted accuracy obtained by our classifiers for each of the classes; that is, 'low penetration', 'medium penetration', and 'high penetration'. The first of these three classes is, unsurprisingly, the class having the highest weighted accuracy. Presumably, this is because it is the most homogeneous class composed by the long tail of areas having 'low penetration' of Airbnb and thus the easiest to estimate. The remaining two classes ('medium' and 'high penetration') are characterized by more heterogeneity in terms of their characteristics and therefore more difficult to be correctly distinguished. Even for these more difficult cases, the median weighted accuracy is above 0.6, thus confirming that our chosen features can effectively be used estimate Airbnb's spatial penetration.

Discussion

Limitations

Our approach has four main limitations, which will inform our future research agenda. First, regression and classification analyses cannot determine casual relationships. While we can justifiably argue that Airbnb's spatial penetration can be explained and predicted for the eight cities under study, we cannot establish any causal relationship. As future work, it would be interesting to perform a longitudinal study of Airbnb penetration and observe changes to city neighborhoods.

Second, even though our list of cities aimed at capturing a variety of socio-economic conditions, it is not comprehensive. Future work should replicate this study upon more cities, which are not necessarily in the U.S. and have not been early adopters of Airbnb (as most of our cities were). It would be interesting to test whether the growth dynamics in late-adopting cities is similar to those in early-adopting ones, that is, whether the presence of the creative class would still matter.

Third, our U.S. census data is the latest but is almost six years old, and it would be prudent to replicate this study with more recent census data once it becomes

available, to gain confidence in the temporal validity of our findings. In this work, we could only check for temporal validity by going backwards: we checked the differences that would result in using the 2000 census data (rather than the 2010 one), and these differences were indeed negligible.

Finally, this study analyzes Airbnb spatial penetration from the point of view of 'offerings' only (i.e., number of Airbnb listings per area). We have already performed a similar study of Airbnb 'demand' (i.e., number of properties actually rented per area) on the same eight U.S. cities and we have found very similar results. An interesting future study would be to analyze Airbnb penetration while segmenting *by users' demographics*, to shed light onto the impact of the service on different classes of users (for example, on tourists vs. business travelers).

Theoretical Implications

Developing methods to quantify adoption of new technologies offers researchers the ability to understand to what extent their findings are generalizable, and under what circumstances. Previous theoretical models of technology adoptions in the city context have overly emphasized the importance of a factor – distance from the city center [51, 52, 53]. Yet, we found other factors to be as important, for example the presence of residents who work in the creative industries. New theoretical models of adoption could be designed, as we now have a more comprehensive understanding of which factors matter.

Practical Implications

One of the main findings of this study was the striking consistency of the results across eight U.S. cities of different nature. This consistency suggests that, to a certain degree, our model could be applied to a city that has not been previously analyzed, to identify areas that tend to be under-represented, understand why those areas are so, and plan interventions to improve the situation.

To see how this could be done from a legal standpoint, we consider once again the recent proposal by Stephen Miller of using "transferable sharing rights" to legalize short-term rental markets like those enabled by Airbnb [5]. These rights could be bought and sold on dedicated web sites, with prices adjusted based on, for example, a neighborhood's economic development plan. Based on our analysis, such sharing rights could be allocated so that a socio-economic deprived area would be allowed to have a high number of sharing rights at a low price and, as such, the area's local economy would benefit (e.g., Airbnb guests tend to shop at local shops). At the same time, an excessive number of short-term rentals in the same neighborhood should be avoided, or else its character and ambiance are bound to be compromised. This could be achieved by limiting the number of sharing rights allocated to areas with higher concentration of youngsters and of people who work in the creative industries, since our analysis revealed that hot-spots of Airbnb rentals are invariably linked to them.

Finally, Airbnb, like most sharing economy platforms, is not penetrating all cities at the same time; rather, some cities will be earlier adopters, while others will be late adopters. One could take our methodology one step further and repeatedly apply it over time, for example on a yearly basis, to identify what factors matter

the most in explaining Airbnb penetration at a given point in time (as previously done for the city of London, UK [13]). When considering an $(N+1)^{th}$ city, one could first identify where it temporally stands in terms of Airbnb penetration trajectory, and then extrapolate from there.

Conclusion

This is the first time that Airbnb's adoption has been analyzed for a wide range of different U.S. cities. We have extracted a variety of geographic, economic, and socio-demographic indicators and shown that, despite the 8 U.S. cities analyzed being rather different in terms of ethnic composition and socio-economic characteristics, in most of them central areas with a strong presence of educated and creative people are those with highest Airbnb penetration. Finally, we have presented a generic prediction model for forecasting Airbnb penetration by exploiting a variety of indicators that also captures non-traditional socio-demographic dimensions such as the presence of creative workers in these areas. We have shown that the proposed model can effectively be used to predict Airbnb penetration in the U.S. cities considered in this study.

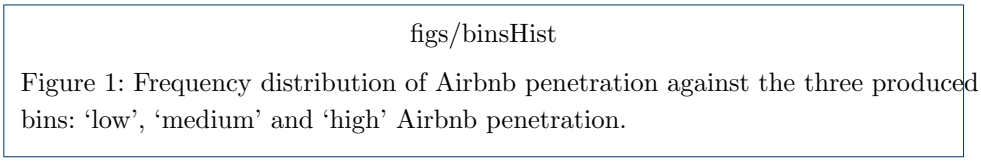

Figure 1: Frequency distribution of Airbnb penetration against the three produced bins: 'low', 'medium' and 'high' Airbnb penetration.

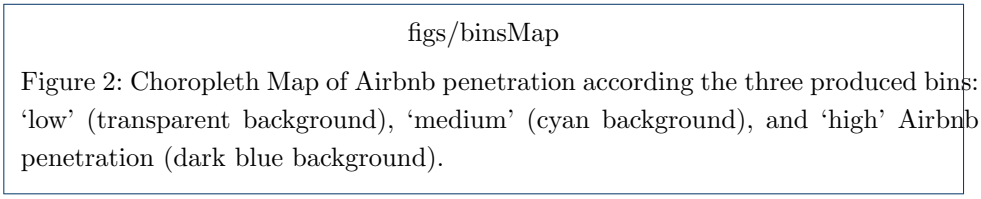

Figure 2: Choropleth Map of Airbnb penetration according the three produced bins: 'low' (transparent background), 'medium' (cyan background), and 'high' Airbnb penetration (dark blue background).

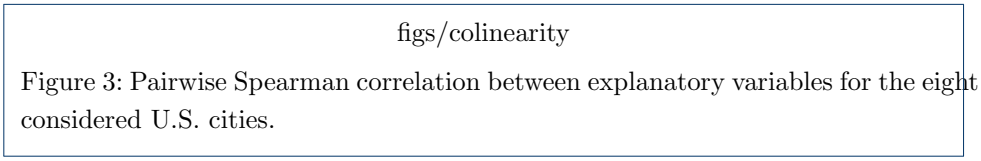

Figure 3: Pairwise Spearman correlation between explanatory variables for the eight considered U.S. cities.

List of abbreviations
U.K.: United Kingdom
U.S.: United States of America
RQ1: Research Question 1
RQ2: Research Question 2
bnb_penetration: Airbnb penetration
distance: Distance to Center
poi: Points of Interest
hotel: Number of Hotels
bus: Bus Stops
popDens: Population Density
race_div: Race Diversity Index
income_div: Income Diversity Index
bohemian: Bohemian Index
talent: Talent Index

figs/censusTempAutocorr

Figure 4: Census 2010 against Census 2000.
Spearman correlation between the explanatory variables derived by the U.S. census released in 2000 against the values they assume when the U.S. census released in 2010 is used. Cells marked by "X" denote correlations that are not statistically significant (p-values greater than or equal to 0.01).

figs/waccAvg

Figure 5: Averaged weighted accuracy of our classifier across the three classes vs. that obtained by the benchmark classifier.

young: Proportion of Young People
unemployment: Unemployment Ratio
poverty: Poverty by Income Percentage
income: Median Household Income
household_value: Median Household Value
owner: Proportion of Owner Occupied Residences
OLS: Ordinary Least Squares
VIF: Variance Inflation Factors
SVM: Support Vector Machines
TP: True Positive
FP: False Positive
TN: True Negative
FN: False Negative

Availability of data and materials
All relevant data will be held in a public repository after acceptance. All relevant data will be available without any restriction.

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

Funding
This work was partially supported by the EPSRC grant EP/L018829/2.
M.M. was partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

Author's contributions
G.Q., D.Q, and L.C. devised the project and the main conceptual ideas. A.G. and G.Q. worked out almost all of the technical details, performed the experiments, and analyzed the data. All authors aided in interpreting the results and contributed to the final manuscript.

Acknowledgements
Not applicable.

Authors' information
G.Q. is a Lecturer in Computer Science at Middlesex University in London, U.K.. His research interests largely fall at the intersection of data mining, web science, social computing, and urban computing.
A.G. is a graduate student in the Department of Computer Science, University College London. His research focuses on spatio-temporal analysis of urban phenomena.
D.Q. is a computer scientist at Bell Labs Cambridge, U.K.. His research area is urban computing.
L.C. is Professor of Pervasive Computing in the Department of Computer Science, University College London. Her research interests include computer-supported collaborative work, social computing, and urban informatics.
M.M. is a Reader in Data Science at the Department of Geography, University College London. His research interests include data mining&modeling of behavioral and social datasets and applied machine learning.

Author details
[1]Middlesex University, London, UK. [2]University College London, London, UK. [3]Bell Labs, Cambridge, UK.

References
1. Kaplan, R.A., Nadler, M.L.: Airbnb: A case study in occupancy regulation and taxation. U. Chi. L. Rev. Dialogue **82**, 103 (2015)
2. Sundararajan, A.: What airbnb gets about culture that uber doesn't. Harvard Business Review **11** (2014)
3. Gurran, N., Phibbs, P.: When tourists move in: how should urban planners respond to airbnb? Journal of the American planning association **83**(1), 80–92 (2017)

figs/wacc

Figure 6: Weighted accuracy of our classifier.

4. Airbnb. About Us. https://www.airbnb.co.uk/about/about-us (cited August 2017)
5. Miller, S.R.: First principles for regulating the sharing economy. Available at SSRN (2015)
6. Florida, R.: Bohemia and economic geography. Journal of Economic Geography **2**(1), 55–71 (2002)
7. Florida, R.: The Rise of the Creative Class–Revisited: Revised and Expanded. Basic books (2014)
8. Thebault-Spieker, J., Terveen, L., Hecht, B.: Toward a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit. ACM Transactions on Computer-Human Interaction **24**(3), 21–12140 (2017). doi:10.1145/3058499
9. Hughes, R., MacKenzie, D.: Transportation network company wait times in Greater Seattle, and relationship to socioeconomic indicators. Journal of Transport Geography **56**, 36–44 (2016). doi:10.1016/j.jtrangeo.2016.08.014
10. Stark, J., Diakopoulos, N.: Uber seems to offer better service in areas with more white people. That raises some tough questions. https://goo.gl/sJGTSt (cited March 2016)
11. Zervas, G., Proserpio, D., Byers, J.: The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. Boston U. School of Management Research Paper (2013-16) (2016)
12. Varma, A., Jukic, N., Pestek, A., Shultz, C.J., Nestorov, S.: Airbnb: Exciting innovation or passing fad? Tourism Management Perspectives **20**, 228–237 (2016)
13. Quattrone, G., Proserpio, D., Quercia, D., Capra, L., Musolesi, M.: Who benefits from the sharing economy of airbnb? In: Proceedings of the 25th International Conference on World Wide Web (WWW '16), pp. 1385–1394 (2016)
14. Edelman, B.G., Geradin, D.: Efficiencies and Regulatory Shortcuts: How Should We Regulate Companies like Airbnb and Uber? Harvard Business School NOM Unit Working Paper (16-026) (2015)
15. Koopman, C., Mitchell, M., Thierer, A.: The Sharing Economy and Consumer Protection Regulation: The Case for Policy Change. Journal of Business, Entrepreneurship and the Law **8**(529) (2015)
16. Einav, L., Farronato, C., Levin, J.: Peer-to-peer markets. Technical report, National Bureau of Economic Research (2015)
17. Ranchordás, S.: Does Sharing Mean Caring: Regulating Innovation in the Sharing Economy. Minnesota Journal of Law, Science and Technology **16**, 413 (2015)
18. Cohen, M., Sundararajan, A.: Self-regulation and innovation in the peer-to-peer sharing economy. University of Chicago Law Review Dialogue **82**, 116 (2015)
19. Zale, K.: Sharing property. University of Colorado Law Review **87**, 501 (2016)
20. Ikkala, T., Lampinen, A.: Monetizing network hospitality: Hospitality and sociability in the context of airbnb. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15), pp. 1033–1044 (2015). ACM
21. Malhotra, A., Van Alstyne, M.: The dark side of the sharing economy. . . and how to lighten it. Communications of the ACM **57**(11), 24–27 (2014)
22. Interian, J.: Up in the Air: Harmonizing the Sharing Economy through Airbnb Regulations. Boston College International and Comparative Law Review **39**, 129 (2016)
23. Numbeo. Cost of Living. https://www.numbeo.com/cost-of-living/ (cited August 2017)
24. Brunn, S.D., Williams, J., Zeigler, D.J.: Cities of the world: world regional urban development. Rowman & Littlefield (2003)
25. Airbnb. Economic Impact. http://blog.atairbnb.com/economic-impact-airbnb/ (cited August 2017)
26. Jacobs, J.: The life and death of great American cities. New York: Random House (1961)
27. Whyte, W.H.: City: Rediscovering the center. University of Pennsylvania Press (2012)
28. Gehl, J.: Cities for people. Island press (2013)
29. Fraiberger, S.P., Sundararajan, A.: Peer-to-peer rental markets in the sharing economy. NYU Stern School of Business Research Paper (2015)
30. Meyer, P., McIntosh, S.: The usa today index of ethnic diversity. International Journal of Public Opinion Research **4**(1), 56–58 (1992)
31. Jost, L.: Entropy and Diversity. Oikos **113**(2), 363–375 (2006)
32. Clifton, N.: The "creative class" in the uk: an initial analysis. Geografiska Annaler: Series B, Human Geography **90**(1), 63–82 (2008)
33. Florida, R.: The economic geography of talent. Annals of the Association of American Geographers **92**(4), 743–755 (2002)
34. The Wall Street Journal. Airbnb to Add More Services to Home-Sharing Business. https://www.wsj.com/articles/airbnb-to-add-more-services-to-homesharing-business-1390929383 (cited August 2017)
35. Tech Insider. Two-Thirds Of Airbnb's Hosts Don't Have Full-Time Jobs. http://www.businessinsider.com/airbnb-new-services-food-and-cleaning-2014-1?IR=T (cited August 2017)
36. Zweig, M.: What's class got to do with it?: American society in the twenty-first century. Cornell University Press (2004)
37. Census.gov. Poverty Thresholds. https://www.census.gov/cps/data/povthresholds.html (cited August 2017)
38. Census.gov. Geographic Terms and Concepts - Census Tract.

https://www.census.gov/geo/reference/gtc/gtc_ct.html (cited August 2017)
39. Kennedy, P.: A guide to econometrics. MIT press (2003)
40. Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W.: Applied linear statistical models. Irwin Chicago (1996)
41. Hair Jr, J.F., Anderson, R.E., Tatham, R.L., William, C.: Black (1995), multivariate data analysis with readings. New Jersy: Prentice Hall (1995)
42. Marquaridt, D.W.: Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics **12**(3), 591–612 (1970)
43. Tobler, W.R.: A computer movie simulating urban growth in the detroit region. Economic geography **46**(sup1), 234–240 (1970)
44. Legendre, P.: Spatial autocorrelation: trouble or new paradigm? Ecology **74**(6), 1659–1673 (1993)
45. F Dormann, C., M McPherson, J., B Araújo, M., Bivand, R., Bolliger, J., Carl, G., G Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., *et al.*: Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography **30**(5), 609–628 (2007)
46. Milojević, S.: Power law distributions in information science: Making the case for logarithmic binning. Journal of the American Society for Information Science and Technology **61**(12), 2417–2425 (2010)
47. Sun, Y., Wong, A.K., Kamel, M.S.: Classification of imbalanced data: a review. International Journal of Pattern Recognition and Artificial Intelligence **23**(04), 687–719 (2009)
48. Kotsiantis, S., Kanellopoulos, D., Pintelas, P., *et al.*: Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering **30**(1), 25–36 (2006)
49. Pan, Y., Jackson, R.T.: Ethnic difference in the relationship between acute inflammation and serum ferritin in us adult males. Epidemiology & Infection **136**(3), 421–431 (2008)
50. Rogerson, P.: Statistical methods for geography. Sage (2001)
51. Hecht, B., Stephens, M.: A Tale of Cities: Urban Biases in Volunteered Geographic Information. In: Proceedings of the 8th International AAAI Conference on Welogs and Social Media (ICWSM 2014), pp. 197–205 (2014)
52. Ishida, K.: Geographical bias on social media and geo-local contents system with mobile devices. In: System Science (HICSS), 2012 45th Hawaii International Conference On, pp. 1790–1796 (2012). IEEE
53. Quattrone, G., Capra, L., Meo, P.D.: There's No Such Thing as the Perfect Map: Quantifying Bias in Spatial Crowd-sourcing Datasets. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15), pp. 1021–1032 (2015). ACM