Location Choice, Portfolio Choice*

Ioannis Branikas† Harrison Hong‡ Jiangmin Xu§

First Draft: September 2016

This Draft: October 2018

Abstract

Households hold undiversified stock portfolios of firms headquartered near their city of residence. Leading explanations like the familiarity heuristic assign a causal role for proximity. The literature neglects that distance is endogenous: households may locate based on unobservables such as optimism about a city's economic prospects, which can be correlated with latent local-stock demand. We propose as instruments that older households closer to retirement prefer to locate in areas with recreation and mild climate for non-pecuniary reasons. Causal estimates are significant but much smaller than those in the literature. Location choices have a larger impact on household portfolios than proximity.

*We thank Ulrich Mueller, Mark Watson, Chris Sims, Bo Honore, Matti Keloharju, Selale Tuzel, Baolian Wang, Kirill Evdokimov, Motohiro Yogo, Atif Mian, Jakub Kastl, Jeffrey Kubik and participants at Econometrics and Finance seminars at Princeton University, University of Toronto, Johns Hopkins University, Columbia University, University of Maryland, University of Oxford, INSEAD, the 2016 LACEA/LAMES Conference, the 2016 China Five-Star Workshop in Finance, the 2016 NYU Shanghai Volatility Institute Conference, the 2017 WFA Conference, the 2017 CICF Conference, and the 2017 Helsinki Finance Summit for helpful comments.

†University of Oregon

‡Columbia University and NBER

§Peking University

1. Introduction

A long-standing puzzle in financial economics is that households hold undiversified stock portfolios tilted toward firms headquartered near where they reside. Contrary to the market portfolio prescription of the CAPM (Sharpe (1964)), households load on local stocks regardless of their market value. In canonical regressions of household stock-portfolio weights on demographic and stock characteristics, *distance* from household residence to firm headquarters emerges as a key explanatory variable. This local-bias appears in many countries.[1] This phenomenon is a granular and more puzzling version of the international home-bias puzzle, where households in different countries tilt toward stocks in their own country (French and Poterba (1991)). In the international setting, portfolio costs or restrictions at least seem plausible impediments toward diversification.

Given the potentially high costs of under-diversification for households, many theories have been given for this local bias in the literature.[2] The leading interpretation of local bias is a familiarity heuristic (e.g., Heath and Tversky (1991), French and Poterba (1991), Huberman (2001)), whereby investors favor local stocks that they view as having competence—be it the company they work for, companies near them that they know friends at, or even the telephone company that services their homes. The most recent studies using the best practice portfolio-return methodologies find no evidence that local stock picks of households out-perform their distant stock picks, thereby pointing to the familiarity heuristic as the main causal mechanism as opposed to an alternative explanation having to do with rational information acquisition (see, e.g., Seasholes and Zhu (2010)).[3]

[1] Prominent studies include the US (Zhu (2002)), Finland (Grinblatt and Keloharju (2001)) and China (Feng and Seasholes (2008)) to name a few.

[2] There is a sizeable literature examining the potential costs of under-diversification in stock portfolios and financial mistakes or literacy more generally (see, e.g., Campbell (2006), Bayer, Bernheim, and Scholz (2009), Agarwal, Driscoll, Gabaix, and Laibson (2009), Lusardi and Mitchell (2007)). Many households around the world have concentrated local stock holdings and little diversification through other investment vehicles (see, e.g., Keloharju, Knupfer, and Rantapuska (2012)). The costs of foregone diversification would seem to be large unless their local stock picks can significantly outperform the market.

[3] This stands in contrast to professional investors, where Coval and Moskowitz (2001) find such an informational advantage in the local trades of mutual funds managers. Another causal explanation for local bias of households is Keeping up with the Joneses' preferences (Luttmer (2005), Charles, Hurst, and Roussanov

We point out in this paper that extant empirical work on local bias subtly but crucially assumes that households locate randomly, which is likely to be counterfactual. Notably, in endogenous location choice models from urban economics, agents optimally locate in cities that provide them with the highest utility (e.g., Bajari and Kahn (2005), Bayer, Ferreira, and McMillan (2007)). This sorting or self-selection depends on both pecuniary (i.e., productivity) and non-pecuniary (i.e., life, leisure or recreational) household motives. While some of these factors are observable to the econometrician, such as older households prefer a city with a mild climate or recreation (i.e., a non-pecuniary motive) or families prefer a city with affordable housing (i.e., a pecuniary motive), many others are unobservable to the econometrician (i.e., latent factors). One of the most important latent factors we argue is subjective expectations about the economic prospects of a city. Households that are identical in every observable dimension locate in different cities if they hold heterogenous enough subjective expectations. Such optimal spatial-sorting models are consistent with migration patterns that one sees in the US (e.g., Clark and Hunter (1992), Bishop (2007), Kennan and Walker (2011), Kaplan and Schulhofer-Wohl (2012), Diamond (2016)).

Even though empirical studies control for detailed observable outcomes such as occupational status of the household or professional proximity, they cannot account for latent subjective expectations about the economic prospects of different cities. Households naturally prefer to move to areas which they view as having a bright economic future not only for themselves but for their family or future generations. But these latent expectations in their location choices are likely to be naturally correlated down the line with optimism about local versus distant stocks to the extent stocks are sensitive to economic conditions of the region of their firm headquarters.

More importantly, their latent expectations about cities that they did not move to (as we demonstrate below) play an important role in their portfolio choices as well. In other words, to what extent does proximity play a causal role in local bias and to what extent

(2009)) which leads to a demand for local stocks as a form of hedging (DeMarzo, Kaniel, and Kremer (2004), Gomez, Priestley, and Zapatero (2009), Hong, Jiang, Wang, and Zhao (2014)).

does it simply reflect selection bias? As a thought experiment, if we were to randomly locate households in different cities, would they still exhibit the same degree of local bias? The location selection bias is ultimately an omitted-variables problem, whereby unobservable location factors correlated with investment-demand shocks are ignored, violating the strict exogeneity assumption on distance in a standard portfolio weights regression.

As such, we develop a methodology to account for the effect of endogenous location decisions on household portfolio choice. We consider two widely used reduced-form portfolio specifications: a linear model where the dependent variable is household portfolio deviations from a market benchmark and a non-linear Tobit model where the dependent variable is portfolio weights. In either setting, distance between a household's MSA and the MSA of the firm's headquarters is the independent variable of interest. This variable is endogenous as we have pointed out to the extent households are choosing a city based on latent expectations about future city prospects.

To quantify the effect of our endogenous location-decision adjustment on the local bias of portfolio choices, we augment standard city demographics by first hand collecting new data on a city's amenities using the ratings of Places Rated Almanac, which is a perennial best-selling guide going back to 1981 for families figuring out where to locate. The MSA features on which we focus are income per capita, unemployment, home price index, population, transportation, colleges, healthcare, crime, recreation and climate.

Our sample of household portfolios comes from a US brokerage database with roughly 9,000 households living in 57 MSAs with a population above 750K, during the period of 1991-1996 (Odean (1999), Barber and Odean (2000)). This sample, in which high income households have a significant fraction of their assets in stocks, is widely-used in the local-bias literature and hence allows us to demonstrate the importance of locational sorting. The data has a variety of household demographics such as age, gender, and family size, which is also key for our analysis. Moreover, the local bias in this earlier sample is remarkably similar to the local bias documented in the most recent brokerage house sample (Gargano and Rossi

(2018)). As such, our analysis is likely to apply to current and future studies.

The instruments that we propose is that older households, whose heads are presumably closer to retirement, prefer to live in areas with mild year-round climates (e.g., the Portland as opposed to the Milwaukee metropolitan area) or areas with recreation like golf courses, tennis courts or outdoor pools (e.g., the Miami metropolitan area as opposed to the Dallas-Fort Worth metroplex). A location choice model, on which we base the first-stage regression of distance, finds that the interactions of the age of the household with the recreation and climate scores of a MSA are highly statistically significant. In fact, in terms of impact, these two variables are among the strongest predictors of household location choice in our sample. These motives for location are non-pecuniary.

Following the guidelines of Roberts and Whited (2013), we conduct balance tests to show that other observable MSA features which are likely to be associated with pecuniary motives, such as the income per capita, the unemployment rate and even the financial characteristics of the local stocks, cannot be not predicted by the climate or recreation score in a statistically significant way.

For our two instruments to be valid, they also have to be uncorrelated with unobservables in the second-stage portfolio weights regression. Our exclusion restriction is that older households do not have different subjective expectations relative to younger households about stocks headquartered in the Miami metropolitan area as opposed to stocks headquartered in the Metroplex - two MSAs with very different recreation scores but similar income per capita, population density and local stocks - for any reason other than the geographical proximity.

This restriction is plausible. The large literature on household finance and asset pricing emphasizes that household demographics (e.g., age) might influence beliefs and that city features (e.g., income per capita) might affect the expected returns of local stocks. Therefore, we always control (in both stages) for household demographics (including age) - to difference away MSA-invariant beliefs - and demographics of the MSAs of the stocks' headquarters

(including the climate and recreation score) - to difference away household-invariant expectations about the cities. But the interaction of age and MSA recreation and climate ought to be excluded.

To correct the endogeneity of distance in the linear deviations model, we can use standard instrumental-variables regression approaches. As the investment universe, we consider stocks that belong to the Russell 1000 Index, an index that includes the largest 1000 stocks based on market capitalization.[4] In a variety of specifications, we estimate a substantially decreased causal effect of distance on the household stock-portfolio weight — around 30% lower than the OLS estimates. But the causal estimate remains economically and statistically significant. We also propose several other instruments using other household demographics interacted with climate and recreation as a robustness exercise.

To correct the non-linear portfolio model, we use a control-function approach. Our optimal location choice model allows us to recover the expected location utility of a household in a city and hence the probability that it locates there. Similar to Heckman (1977), these location probabilities can then be added in the Tobit weights regression as extra covariates that capture unobserved locational shocks. To the extent that there is no location selection bias, introducing these probabilities should not affect the estimate of the coefficient on distance. We get a larger correction correction of 43% of the distance coefficient, compared to around 30% for the linear model.

These estimates suggest a structural model of household portfolio choice and local bias depending on household priors about stocks in a city - which we can proxy using Pearson residuals from our location choice model - and familiarity - which we model as investor confidence about stock-payoff signals increasing with proximity to the stock headquarter. Households are risk neutral but face short-sales constraints and quadratic trading costs. These assumptions are realistic for our investors since they hold concentrated positions, do not short and face high trading costs. Households form their portfolios based on both

[4]In the online Appendix, we show that our results apply also for the extended investment universe of Russell 3000.

priors about stocks in a city and also i.i.d. private signals about individual stock payoffs. Thus, our model provides two rationales for local bias: optimistic priors and familiarity. We estimate our model and gauge the relative importance of these two factors. We find that the household priors account for 60% of local bias and familiarity or distance 40% using goodness-of-fit tests.

2. Data

2.1. MSA Demographics

Following previous work on local bias in the US (e.g., Coval and Moskowitz (1999)), we exclude MSAs in Alaska, Hawaii and Puerto Rico. Our main analysis features 57 MSAs with a population of at least 750,000 at the end of 1996. We apply this filter only for tractability and to make sure that the number of broker's investors in each MSA is high enough to estimate their location probabilities precisely.[5]

The traditional list of variables that urban studies have used consists of the total income drawn from the Bureau of Economic Analysis (BEA), the unemployment rate extracted from the Bureau of Labor Statistics (BLS), and the house price index (HPI) taken from the Federal Housing Finance Agency (FHFA). These last variables are observed at an annual frequency.

In addition we contribute to the literature of location choice by collecting additional variables that can capture many aspects of a city that are bound to be very relevant when it comes to moving into a given MSA. The data on these MSA livability scores are extracted from the 1993 edition of *Places Rated Almanac* (by Savageau and Boyer (1993)). The almanac contains ratings with respect to (i) the ability to meet transportation needs, (ii) college opportunities, (iii) the supply of health care, (iv) crime, (vi) the supply of recreational

[5]In the Online Appendix, we repeat the analysis for 80 MSAs with a population of at least 500K in the beginning of 1991 and obtain similar results.

assets, and (vii) climate mildness. The analytical definition of these variables is given in the Appendix. The higher the score of transportation, health care, recreation or climate, the better the living conditions in terms of these variables. On the other hand, a high score of crime in a MSA indicates more danger.

We are particularly interested in the climate and recreation scores. The correlation between the two is roughly 0.3, since the recreation score is based on the access to outdoor activities, which require some temperate weather for at least part of the year. But they are not identical. The Miami metropolitan area scores high in recreation, since it features lots of golf courses and tennis courts, but not in climate, since it can be very hot and humid in the summer. In contrast, the Portland metropolitan area scores high in climate for having mild weather year round, but not in recreation, since it does not have a lot of golf courses.

We present the summary statistics of the MSA demographics in Panels A of Table 1.[6] The top 10 MSAs in terms of climate and recreation are those known as retirement destinations and include cities in Florida, Arizona and Southern California (e.g., San Diego).

2.2. Household Demographics

Our household investment data are drawn from the database of a national discount brokerage firm. See Barber and Odean (2000) for detailed descriptions. The dataset is an unbalanced panel of month-end account statements from approximately 78,000 households at the stock level (CUSIP). The sample period spans from January 1991 to November 1996. Most households have multiple accounts which we aggregate, in order to obtain their total long positions in a given stock. As Ivković and Weisbenner (2005) report, the majority of accounts is non-retirement (e.g., cash or investment) and the few retirement accounts do not refer to 401(k) plans. Therefore, mechanical effects on the stock choice from the shares of an employer are limited.

[6]The mean income per capita is 21.6 thousand dollars, while the mean unemployment rate is 6.25 percentage points. The mean HPI is 94.8. The mean population is 2.5 million with a standard deviation of 2.7 million.

In our analysis, we omit households whose demographic information is incomplete. Specifically, we require households to have a non-missing address ZIP-Code, income, family size, age, gender and marital status of the head. This criterion decreases the sample size to approximately 40,000. We also require the observability of the job code of a household's head, according to which its occupation is classified as (i) professional or technical, (ii) administrative or managerial, (iii) sales or service, (iv) white-collar or (v) blue-collar. That criterion further reduces the number of household to around 17,500.

Unfortunately, the data of the discount broker do not contain any information about the education, race and industrial sector of the household's head. To correct for that, we follow Korniotis and Kumar (2011) and extract from Census 1990 the education status (i.e., the probability of holding a B.A. or higher degree) and the racial profile (i.e., the probability of being White, Black, Hispanic, Asian or other) at the household's ZIP-Code. Moreover, we use the distribution of the employed persons into industries at the ZIP-Code level to measure the household's (expected) professional industrial proximity to a stock, in the style of Massa and Simonov (2006).[7] Requiring the complete observability of these additional household demographics leaves with 12,892 households.

By focusing on the selected 57 MSAs, we derive our final sample, which consists of 8,688 unique households with complete information on demographics and stock portfolios.[8] These households do not move across MSAs, but stay in their original location either until the last date in the data or until they close their accounts.[9] Their first time-series observations comprise the sample of our location choice model. The summary statistics of the household demographics are presented in Panel B of Table 1.[10]

[7]For example, New Yorkers living in Upper East Side are expected to be familiar with stocks in the financial sector, since many investment bankers reside there.

[8]In the Online-Appendix, when we focus on 80 MSAs, we have 10,261 households.

[9]According to the US Census Bureau, the average percentage of movers during our sample period (1991-1996) was on average about 17% (http://www.census.gov/newsroom/press-releases/2015/cb15-47.html). This means that, roughly, a household would be expected to change residence every 6 ($\approx 1/0.17$) years. Given that our own sample period is six years, we expect that only few households in the data moved.

[10]The income of households in our sample has a mean of 101.85 thousand dollars and a median of 87.5 thousand dollars. This is to be expected since only households with sufficient income would participate in the stock market to begin with. The mean age is 52 years. About 56% of the households are professionals and

2.3. Stock Financial Characteristics

The universe of stocks that we study in our main analysis consists of stocks that were ever members of the Russell 1000 Index during the sample period. We focus only on stocks located in the same 57 MSAs as above, with a complete list of financial characteristics as described below.[11] This filter leads us to a total number of 1,193 different stocks for the whole period.

Monthly data on stock prices and returns are drawn from CRSP, while firm accounting variables are collected from Compustat at a quarterly frequency. Stocks with missing observations are dropped from the sample. The list of financial variables that we use consists of the price, the market capitalization (Size), the book-to-market ratio (BTM), the turnover ratio (i.e., Turnover, defined as volume over number of shares outstanding), the momentum (i.e., Momentum, defined as the past annual return), the volatility (i.e., Volatility, defined as the standard deviation of monthly returns in the past year), profitability (i.e., Profitability, defined, as in Novy-Marx (2013), as the ratio of past annual gross profits to assets) and the investment (i.e., Investment, defined as the past annual growth rate of assets).

All the above variables are constructed at a monthly frequency. We assume that a household's investment decision in month t is based on the stocks' price in that month and the above risk factors in month $t-1$.[12] We also use the Fama-French industry classification of stocks into 17 categories based on the four-digit SIC code, which is available from Kenneth

27% are managerial. Sales service, white collar and blue collar accounts comprise about 8%, 5% and 4% of the total respectively. Approximately, 92% of the households are headed by a male and 73% of the heads are married. The average family size is 2.5. At the household ZIP-Code level, the expected professional industrial proximity is 8%, while the expected advanced educational attainment is 36%. The average percentage of Whites is 81%, Blacks 6%, Hispanics 7% and Asians or of other race 6%.

[11] Since Compustat contains only the most recent headquarters' addresses of the stocks, a variety of sources (e.g., EDGAR, COMPHIST, Who Owns Whom, etc.) is utilized to ensure that the headquarter information in the sample period is accurate. Pirinsky and Wang (2006) identify 118 firm relocations from 1992 to 1997 and Tuzel and Zhang (2017) about 300 from 1990 to 2005. In most cases, the firms that moved were small and not members of Russell 1000.

[12] As in Fama and French (1992), to make sure that firms' balance sheet information is known to investors, we match the accounting variables from the fiscal year $t-1$ with the stock prices from the July of year t until the June of year $t+1$.

R. French's website.[13] We depict the summary statistics of the stock financial characteristics in Panel E of Table 1.[14]

2.4. Geographical Distribution of Household Stock Holdings

Using the US Census Bureau geographical coordinates of the ZIP-code of every household and the ZIP-code of the headquarters of every stock, we calculate their spherical distances, which are the key variable in our study.[15] The geographical distribution of the 8,688 households in our sample and the 1,193 stocks in Russell 1000 is presented in Figure 1 via a map of latitude and longitude coordinates of the households and the stocks' headquarters. Overall, the sample is dispersed enough to be representative of the US population. In terms of the potential local bias, households are always located near the headquarters of some firms.

The portfolio positions of households are summarized in Panel C of Table 1. The mean value of a household's portfolio in common stocks is about \$31K (averaged across time periods from 1991 to 1996), while the median value is about \$11K.[16] The standard deviation of the household stock holdings' value in our sample is \$126K. To assess the trading activity of the selected households, we calculate their sales and purchase turnover as Barber and Odean (2000). On average, the monthly sales turnover is 3.20%, while the monthly purchase turnover is 4.05%. In other words, the retail investors in the sample are not passive, since they buy 48.6% and sell 38.4% of their portfolio every year.[17]

[13] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

[14] The mean market capitalization of a stock in the Russell 1000 index is around 3.6 billion dollars. The mean book-to-market ratio is 0.56. The mean monthly turnover ratio is 10%. The mean past 12-month return is 12% and the mean monthly volatility is 9%. The mean profitability is 34%, while the mean investment is 20%. The industrial composition of the Russel 1000 Index is reflected by the 17 Fama-French industry classification; 25% of the stocks belong to the "Other" industry category, 18% of them are in "Finance" (referring to banks, insurance companies and other financials), while 10% belong to the "Machines" category (for machinery and business equipment).

[15] We measure distance in degrees. Multiplying by $2\pi R/360$ converts it to miles (kilometers), where $R \approx$3,963 miles (6,378 kilometers).

[16] The report by the US Census Bureau on net worth and asset ownership of households in 1998 and 2000 shows that in 1998, the median value of holdings in stocks for a typical US household is \$16,800 (https://www.census.gov/prod/2003pubs/p70-88.pdf). This information indicates that our sample is similar to the stock holding situation of US households in the '90s.

[17] For all households and all stocks in the database, Barber and Odean (2000) document an average annual

Furthermore, on average, a household in our sample has a portfolio weight of 10.31 bps on a Russell 1000 stock and holds 2.32 stocks. The standard deviation of the number of stocks is 2.27, indicating that most of the households in the sample are under-diversified, even as their stock holdings comprise a substantial fraction of their assets. The median number of stocks that a household holds is 1.7, while the standard deviation of a portfolio weight is 0.03.[18]

In Panel D, we report the mean distance of a household's residence to a Russell 1000 firm headquarters, which is 17.6 degrees. The standard deviation is 11.73. These figures can be compared to the distance between California and New York, which is approximately 57 degrees. In addition, we also construct dummies indicating whether a stock is headquartered more than 250 miles and 100 miles away from a household's residence. The average percentage of households that are away from a stock's headquarters according to these metrics is about 90%.

2.5. Summary Statistics on Local Bias of Stock Portfolios

The summary statistics of the local bias (LB) in our household stock holdings data are given in Table 2 and are constructed as in Coval and Moskowitz (1999). Column 2 (labeled "Avg. Distance from Holdings") reports the average *portfolio weighted distance* of households from their stock holdings, defined as $\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} w_{i,j} dist_{i,j}$, where $dist_{i,j}$ is the ZIP-code distance between a household i's residential area and the headquarters area of a stock j, $w_{i,j}$ is the household i's portfolio weight on stock j, I is the total number of households and J is the total number of stocks in the investment universe. Column 3 (labeled "Avg. Distance from Benchmark") reports the average portfolio weighted distance of households from the Russell 1000 benchmark portfolio, computed as $\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} \bar{w}_j dist_{i,j}$, where $\bar{w}_j$ is the Russell 1000 benchmark portfolio weight on stock j. Row 1 has as benchmark the equally weighted

portfolio turnover of 75%.

[18] These figures are very similar to the portfolio summary statistics reported by Gargano and Rossi (2018) for a recent broker during the period January 2013 to June 2014.

portfolio, while Row 2 refers to the value-weighted portfolio. Column 5 (labeled "Difference") reports the average difference between Column 3 and Column 4, which is essentially the average local bias of households in distance units. Column 6 (labeled "% Bias (LB)") reports the local bias (LB) measure as a percentage. Column 7 reports the t-statistics for the LB measure. Independent of which benchmark is used (the values are about the same), the local bias is always high in terms of both magnitude and statistical significance. Specifically, using the equally weighted portfolio, the local bias is 8.29 or 45.45%, while, using the value-weighted portfolio, it is slightly decreased to 8.26 or 43.72%.[19]

3. Accounting for Location Choice in Reduced-Form Portfolio-Choice Regressions

In this section, we present a simple framework that highlights the implications of a household's location choice on its subsequent investment decisions. We index households with i, stocks with j, and periods with t; overall, there are T periods in each of which live I_t households that can potentially invest in J stocks. The total number of cities, throughout the years, is C. We denote with c the city in which household i resides, and with h the city in which stock j is headquartered.

3.1. Location Choice

Since in our data households do not move, we only model their location choice in the beginning of their time series. In line with a standard discrete choice model, we decompose the utility that household i derives from a city $\ell = 1, \ldots, C$ into the sum of an observable component, $V_{i,\ell}$, and an unobservable idiosyncratic shock, $e_{i,\ell}$, and assume that household i is a utility maximizer locating to city c satisfying the following relationship:

[19] The percentage LB is more than four times the local bias that Coval and Moskowitz (1999) report for non-index fund managers in 1995.

$$c = \arg \max_{\ell \in \{1,\ldots,C\}} \{V_{i,\ell} + e_{i,\ell}\} \tag{1}$$

Household i's observable utility from a city ℓ is a linear combination of the city's characteristics at the time at which the location decision is made, which we group into a $K \times 1$ vector $\boldsymbol{z}_\ell$. In our empirical analysis, this vector consists of the city's income per capita, unemployment rate, house price index, population and livability scores for its transportation, colleges, health care, crime, recreation and climate. On the other hand, household i's unobservable utility from city ℓ refers to location factors that we, as econometricians, cannot observe, such as *subjective expectations*.[20]

Although households in a given period view the same city characteristics, they value them differently, i.e.:

$$V_{i,\ell} = \boldsymbol{\rho}_i \boldsymbol{z}_\ell \tag{2}$$

where $\boldsymbol{\rho}_i$ is the vector of household i's responses. In particular, we assume observed heterogeneity in preferences through a matching structure. That is, we decompose $\boldsymbol{\rho}_i$ into a component that is common across all households, $\boldsymbol{\rho}$, and a component that linearly depends on household i's $M \times 1$ vector of demographics, $\boldsymbol{D}_i$, (through a $K \times M$ matrix of parameters $\boldsymbol{\Pi}$), i.e.:

$$\boldsymbol{\rho}_i = \boldsymbol{\rho} + \boldsymbol{\Pi} \boldsymbol{D}_i \tag{3}$$

The vector of household i's demographics, $\boldsymbol{D}_i$, that we use in our empirical analysis has as elements its its family size and the age, gender and marital status of its head. We refer to these variables a "Household Location Demographics".[21] By combining Equations (2) and

[20] Of course, $e_{i,\ell}$ also refers to other factors during the location decision process that are uncorrelated with latent demand for stocks.

[21] We also have readily available data for the income and occupation of the households' heads. However, since these variables could be the outcome of a location decision, we use them only in the portfolio analysis.

(3), we eventually represent household i's observed utility from locating in city ℓ as:

$$V_{i,\ell} = \underbrace{\sum_{k=1}^{K} \rho_k z_{\ell,k}}_{\delta_\ell} + \underbrace{\sum_{k=1}^{K} \sum_{m=1}^{M} \pi_{k,m} D_{i,m} z_{\ell,k}}_{\mu_{i,\ell}} \tag{4}$$

where δ_ℓ is the observed utility from the characteristics of city ℓ that is common for all households, while $\mu_{i,\ell}$ is the observed utility from the characteristics of city ℓ which is different across households. Equation (4) implies that once we estimate the location parameters $\boldsymbol{\theta}^{loc} \equiv (\boldsymbol{\rho}, \boldsymbol{\Pi})$ from the data, we will have also estimated the observed utilities of household i from all the available locations, $\{V_{i,\ell}\}_{\ell=1,\ldots,C}$.

Next, we define household i's maximum order statistic with respect to a city c as:

$$v_{i,c} = \max_{\ell \in \{1,\ldots,C\}/c} \{V_{i,\ell} - V_{i,c} + e_{i,\ell} - e_{i,c}\} \tag{5}$$

so that household i's location rule in Equation (1) can be rewritten as:

$$r_{i,c} = \mathbf{1}\left[v_{i,c} < 0\right] \tag{6}$$

where $r_{i,c}$ denotes household i's decision to reside in city c and $\mathbf{1}[\cdot]$ is an indicator function. Assuming that, conditional on the observables, household i's idiosyncratic shocks, $\{e_{i,\ell,t}\}_{\ell=1}^{C}$, are independently and identically distributed according to the extreme value type I distribution, we can calculate the probability with which it resides in city c as follows:

$$p_{i,c} \equiv \mathbb{P}\left(v_{i,c} < 0 \,\middle|\, \{V_{i,\ell}\}_{\ell=1}^{C}\right) = \frac{\exp\left(V_{i,c}\right)}{\sum_{\ell=1}^{C} \exp\left(V_{i,\ell}\right)} \tag{7}$$

3.2. Reduced-Form Portfolio Choice Regressions

We consider two widely-used regression specifications in the literature, which we will then micro-found in Section 5. Since we estimate portfolio parameters for every period separately (thus allowing for time variation in households' portfolio preferences and expectations), we

omit the period subscript t in the discussion that follows.

For tractability reasons, the literature (e.g., Goetzmann and Kumar (2008), Brandt, Santa-Clara, and Valkanov (2009)) has explicitly focused on the households' portfolio under-diversification, by employing a linear model of *excess* portfolio weights relative to the market. We denote the parameters of this first model $\boldsymbol{\theta}^{dev} \equiv \left(\alpha^{dev}, \boldsymbol{\beta}^{dev}, \boldsymbol{\gamma}^{dev}, \delta^{dev}\right)$ and estimate them by running the following linear regression for every month in the sample period:

$$\frac{w_{i,c,h,j} - w_j^{VW}}{w_j^{VW}} = \alpha^{dev} + \boldsymbol{\beta}^{dev}\boldsymbol{x}_j + \boldsymbol{\gamma}^{dev}\boldsymbol{D}_i + \delta^{dev} dist_{i,c,h,j} + \epsilon^{dev}_{i,c,h,j} \tag{8}$$

The dependent variable in the equation's LHS is the percentage deviation of household i's portfolio weight on stock j from the value-weighted portfolio on that stock. $\boldsymbol{x}_j$ is the vector of stock j's financial characteristics - in particular, its size, book-to-market ratio, turnover, momentum, volatility, profitability, investment and industry code. $\boldsymbol{D}_i$ is the vector of household i's demographics, which is as in its location choice problem, but augmented with its head's income, occupation code, expected education, race and professional industrial proximity to a given stock. Importantly, $dist_{i,c,h,j}$ is the distance between household i's ZIP-code in city c and stock j's headquarters ZIP-code in city h.[22] Lastly, $\epsilon^{dev}_{i,c,h,j}$ is household i's idiosyncratic demand shock for stock j, when the former resides in city c and the latter is headquartered in city h. For instance, it could refer to whether household i thinks highly of stock j because of its board members or products.

The caveat of the above specification is that the many zero portfolio weights on stocksare translated to many 100% negative deviations from the market. Specifically, there is a high degree of sparsity that household portfolios exhibit (e.g., including on average only two stocks in Russell 1000). Therefore, an alternative specification, which is consistent with the fact that households do not short, is to assume that household i, residing in city c, decides how

[22]The linear effect of distance on investing is in the spirit of Coval and Moskowitz (1999). In the On-line Appendix, we repeat our analysis for distance indicator variables (e.g. 250 or 100 miles away) used by Ivković and Weisbenner (2005) and Seasholes and Zhu (2010). We also do the same for the log of distance, which is used in the Scandinavian studies of Grinblatt and Keloharju (2001) and Massa and Simonov (2006).

much to invest in stock j, headquartered in city h, according to a linear factor rule censored at zero:

$$w_{i,c,h,j} = (\alpha + \boldsymbol{\beta}\boldsymbol{x}_j + \boldsymbol{\gamma}\boldsymbol{D}_i + \delta dist_{i,c,h,j} + \epsilon_{i,c,h,j})^+ \tag{9}$$

where $(\cdot)^+ \equiv \max\{\cdot, 0\}$ captures both household i's extensive and intensive margin.

In line with a Tobit model, we assume that, conditional on all observables, the error term is distributed according to the normal distribution. When the households' locational decisions are ignored, the conditional mean of $\epsilon_{i,c,h,j}$ is assumed to be zero. Hence, the portfolio parameters to be estimated from the data are $\boldsymbol{\theta}^{port} \equiv (\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta)$, with δ being the main parameter of interest (i.e. the coefficient on the distance variable).[23]

3.3. The Endogeneity Problem of Distance

Regardless of whether the portfolio choice model is linear or non-linear, there is a fundamental endogeneity problem that has not been addressed by the literature. To see why, note that the distance between household i's ZIP-code in city c and the ZIP-code of stock j in city h where it is headquartered can always be expressed as a function of (i) the distance between household i's ZIP-code and the central ZIP-code of city c in which it resides, $dist_{i,c}$, (ii) the distance between the central ZIP-code of city c in which it resides and the central ZIP-code of city h in which stock j is headquartered, $dist_{c,h}$, and (iii) the distance between the central ZIP-code of city h in which stock j is headquartered and stock j's headquarters ZIP-code, $dist_{h,j}$. In short, denoting $S(\cdot)$ this function, we can write that:

$$dist_{i,c,h,j} = S(dist_{i,c}, dist_{c,h}, dist_{h,j}) \tag{10}$$

The need to control for location choice arises from the fact that the distance between the central ZIP-code of city c in which household i resides and the central ZIP-code of city h in

[23] In the spirit of Petersen (2009), when we estimate the model, we use two-way clustered standard errors at the level of the household and the household's city (a.k.a. MSA).

which stock j is headquartered is the *outcome* of household i's location choice. That is, as long as household i is not randomly assigned to the city where it resides, the location rule in Equation (6) implies that:

$$dist_{c,h} = \sum_{\ell=1}^{C} dist_{\ell,h} r_{i,\ell} \tag{11}$$

where every distance between the central ZIP-code of a city ℓ and the central ZIP-code of city h in which stock j is headquartered, $dist_{\ell,h}$, is multiplied by household i's respective indicator function of its decision to live there, $r_{i,\ell} = \mathbf{1}[v_{i,\ell} < 0]$. Having that in mind, it is very likely that $\epsilon_{i,c,h,j}$, i.e. household i's idiosyncratic investment error when it lives in city c and considers investing in stock j headquartered in city h, is correlated with the idiosyncratic location errors, $\{e_{i,\ell}\}_{\ell=1}^{C}$ - especially $e_{i,c}$ and $e_{i,h}$ - as these are summarized by the maximum order statistic of the city c where household i actually resides, $v_{i,c}$. To show such a potential correlation more clearly, we decompose $\epsilon_{i,c,h,j}$ as follows:

$$\epsilon_{i,c,h,j} = \mathbb{E}\left(\epsilon_{i,c,h,j} \left| v_{i,c} < 0, \{V_{i,\ell}\}_{\ell=1}^{C} \right.\right) + \eta_{i,c,h,j} \tag{12}$$

where $\eta_{i,c,h,j}$ is an idiosyncratic stock-city investment error which, by construction, is independent of household i's location decision to reside in city c. That is, $\eta_{i,c,h,j}$ is mean-zero given all observables. As for the conditional expectation of the original idiosyncratic investment error, $\epsilon_{i,c,h,j}$, given household i's decision to live in city c and the observed location utilities $\{V_{i,\ell}\}_{\ell=1}^{C}$, which are estimated from the location choice model in a first stage, it can be calculated as follows:

$$\begin{aligned} \mathbb{E}\left(\epsilon_{i,c,h,j} \left| v_{i,c} < 0, \{V_{i,\ell}\}_{\ell=1}^{C} \right.\right) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{0} \frac{\epsilon_{i,c,h,j} f\left(\epsilon_{i,c,h,j}, v_{i,c} \left| \{V_{i,\ell}\}_{\ell=1}^{C} \right.\right)}{\mathbb{P}\left(v_{i,c} < 0 \left| \{V_{i,\ell}\}_{\ell=1}^{C} \right.\right)} dv_{i,c} d\epsilon_{i,j} \\ &= \psi_{c,h}\left(\{V_{i,\ell}\}_{\ell=1}^{C}\right) \end{aligned} \tag{13}$$

where $\psi_{c,h}(\cdot)$ is an *unknown* control function whose actual form depends on assumptions regarding the *joint* distribution of $\epsilon_{i,c,h,j}$ and $v_{i,c}$. The value of the control function is in principle non-zero, unless $\epsilon_{i,c,h,j}$ and $v_{i,c}$ are independent.[24] Consequently, based on Equations (10) to (13), the distance variable in the portfolio choice regression, $dist_{i,c,h,j}$, is correlated with the original investment idiosyncratic error, $\epsilon_{i,c,h,j}$, through the control function $\psi_{c,h}(\cdot)$. Any estimation procedure that ignores this correlation is destined to yield biased estimates on the respective coefficients δ and δ^{dev} in Equations (9) and (8).[25]

3.4. Identification Strategies

3.4.1. Instruments

We propose as instruments that older households, which presumably are closer to retirement, prefer to live in areas with mild year-round climates or recreation, i.e. non-pecuniary motives for location. Our first-stage regression of distance uses the location choice model in Equations (1) to (7) to predict where the households in our sample reside based on MSA features, household demographics and interactions of the two. As we show below, the variables $LogClimate \times LogAge$ and $LogRecreation \times LogAge$ are among the strongest predictors of household location choice in our sample.

Following the guidelines of Roberts and Whited (2013), we conduct balance tests to show

[24] In that case, in the numerator of Equation (13), we have that:

$$\begin{aligned}\int_{-\infty}^{+\infty}\int_{-\infty}^{0} \epsilon_{i,c,h,j} f\left(\epsilon_{i,c,}v_{i,c}\left|\left\{V_{i,\ell}\right\}_{\ell=1}^{C}\right.\right) dv_{i,c} d\epsilon_{i,c,h,} &= \int_{-\infty}^{+\infty} \epsilon_{i,c,h,j} f\left(\epsilon_{i,c,h,j}\left|\left\{V_{i,\ell}\right\}_{\ell=1}^{C}\right.\right) d\epsilon_{i,c,h,j} \int_{-\infty}^{0} f\left(v_{i,c}\left|\left\{V_{i,\ell}\right\}_{\ell=1}^{C}\right.\right) dv_{i,c} \\ &= \mathbb{E}\left(\epsilon_{i,c,h,j}\left|\left\{V_{i,\ell}\right\}_{\ell=1}^{C}\right.\right) \mathbb{P}\left(v_{i,c}<0\left|\left\{V_{i,\ell}\right\}_{\ell=1}^{C}\right.\right) \\ &= 0\end{aligned}$$

since the conditional mean of $\epsilon_{i,c,h,j}$, given the observables, is zero.

[25] Since more than one observed location utilities enter the control function, the endogeneity bias on the distance coefficient cannot be *ex ante* assessed. The ability of a household to invest in both local and distant stocks further contribute to this. Intuitively though, under the premise that latent location preferences for an area (here, captured by the control function) are positively correlated with unobservable latent investment preferences for stocks headquartered there, we anticipate a *local bias over-estimation* when location choice is ignored.

that our climate and recreation scores have low correlations with other MSA observables that could be relevant for pecuniary motives. In Table 3, we split the sample of metropolitan areas into high versus low climate or recreation groups (based on the corresponding median score) and then calculate the average demographics in each group. None of the pairwise differences between the two groups is found to be statistically significant.[26] In Table 4, we further show that the financial characteristics of a stock cannot be predicted (in a statistically significant way) by the climate or recreation score in the MSA in which it is headquartered.

The balance of the MSA demographics is important since it allows us to have comparisons of MSAs with high versus low climate or recreation scores, while controlling for their income, population density and financial characteristics of the local stocks. But for our instruments to be valid, they also have to be uncorrelated with unobservables in the second-stage portfolio weights regression. We therefore always include as controls household demographics (e.g., age) and demographics of the MSAs of the stocks' headquarters (e.g., unemployment rate, HPI, climate and recreation scores) to difference away MSA-invariant beliefs and household-invariant expectations about the cities. To be cautious, we also control for all the interactions of the household demographics with the demographics of the MSAs of the stocks' headquarters, except for the ones with the climate and recreation scores.

Our exclusion restriction then boils down to older households not having different subjective expectations relative to younger households about stocks headquartered in Miami-Fort Lauderdale-West Palm Beach as opposed to stocks headquartered in Dallas-Fort Worth-Arlington, two MSAs with very different recreation scores but with similar other observable demographics and local stocks, other than through the proximity effect. That is, the reduced-form IV regression is that we replace distance in explaining portfolio decisions with the interaction of age of the household and the recreation score of the MSA, while controlling for household and MSA demographics. It does not matter if locational decisions can

[26]The highest t-statistic is 1.82 for the paired difference in income per capita between the high versus low climate score groups. But even then, the null hypothesis that the means are equal cannot be rejected at the 5% level of statistical significance.

also be driven by other financial or human capital considerations (e.g., Ortalo-Magné and Prat (2016), Hizmo (2015)). It is only important that we find non-pecuniary motives which predict location choice and can be plausibly excluded from the portfolio choice.

We can also increase the number of instruments by considering other demographic attributes of a household including marital status, number of family members and gender, since they might also help measure whether a household is closer to retirement and desiring these amenities. For instance, even controlling for age, married households with many kids are less likely to be in retirement mode and less likely to prefer these amenities. Therefore, as a robustness check, we also interact these other household demographic attributes with recreation and climate and use them as instruments.

3.4.2. GMM Approach in the Linear Model

In the linear model of Equation (8), we directly examine the effect of the instruments on distance using the following first-stage equation:

$$d_{i,c,h,j} = \iota + \boldsymbol{\kappa}\boldsymbol{x}_j + \boldsymbol{\lambda}\boldsymbol{D_i} + \boldsymbol{\phi}\boldsymbol{D}_i \times \boldsymbol{z}_h + \omega_{i,c,h,j} \tag{14}$$

where the interactions of household i' demographics with the demographics and livability factors in city h where stock j is headquartered are excluded from the second-stage excess portfolio weights regressions. Depending on the number of instruments in hand, we can then run a simple IV or GMM regression for the excess household portfolio weights.

We follow Angrist and Pischke (2008))'s guidelines for (multiple) instrumental variables regressions and report the first-stage F-statistics on the excluded instruments, the just-identified estimates when we only use our most prominent instruments - which involves the interaction of age and a city's climate or recreation scores - and the LIML (as opposed to the 2SLS) estimation results. We get robust answers regardless of the specifications we use.

3.4.3. Control Function Approach in the Non-linear Model

In the non-linear model of Equation (9), we use the instruments' effect on the location probabilities. But we also need to restrict the structure of the control function $\psi_{c,h}$ in Equation (13). We invoke the monotonic relationship between household i's observed location utilities, $\{V_{i,\ell}\}_{\ell=1}^{C}$, and its location probabilities, $\{p_{i,\ell}\}_{\ell=1}^{C}$, which allows us to write that:

$$\psi_{c,h}\left(\{V_{i,\ell}\}_{\ell=1}^{C}\right) = \Psi_{c,h}\left(\{p_{i,\ell}\}_{\ell=1}^{C}\right) \tag{15}$$

Based on Equation (15), we now have a new *unknown* control function, namely $\Psi_{c,h}(\cdot)$, in terms of location probabilities, $\{p_{i,\ell}\}_{\ell=1}^{C}$, which capture the impact of unobservable location factors on subsequent investment decisions, *given* residence choice. Combining Equations (9), (12), (13) and (15) yields that the portfolio weights regression correcting for location choice is:

$$w_{i,c,h,j} = \left(\alpha + \boldsymbol{\beta}\boldsymbol{x}_j + \boldsymbol{\gamma}\boldsymbol{D}_i + \delta dist_{i,c,h,j} + \Psi_{c,h}\left(\{p_{i,\ell}\}_{\ell=1}^{C}\right) + \eta_{i,c,h,j}\right)^{+} \tag{16}$$

4. Estimation

4.1. Location Choice Results

The conditional logit estimation results from four models of location choice are presented in Table 5. The dependent variable is an indicator variable that equals one if a household resides in a specific MSA. Since we have 8,688 unique households living in 57 different MSAs, the sample in the estimation consists of 495,216 observations.

Column 1 contains the estimation results based on a specification with all the MSA demographics except for the climate and recreation scores. In Column 2, we add these two extra MSA features. The pseudo-R^2 increases from 0.113 to 0.119, while the Akaike Information Criterion decreases from 62,270 to 61,931 pointing to a better model (e.g.,

Dubé, Luo, and Fang (2017)).

While in Columns 1 and 2, the location choice model is homogeneous (i.e. Π is set to zero in Equation (3)), in Column 3 and 4, we allow households to match to the MSA demographics based on their own demographics. In Column 3, the location choice model includes again the same MSA demographics as in Column 1. Yet, relative to before, the inclusion of all the pairwise interactions enhances the fit of the model, yielding a pseudo R^2 of 0.116 and a AIC of $62,143$.

In Column 4, we add the MSA climate and recreation scores as in Column 2. Again, we see an improved fit. The pseudo R^2 is further increased to 0.122 and the AIC is further decreased to $61,770$. Moreover, since the model in Column 4 nests the models in Columns 1, 2 and 3, we perform a likelihood ratio test and reject the restricted versions of the model for any reasonable level of statistical significance. Hence, in what follows, to account for the effect of location decisions in the non-linear model of portfolio choice, we use control functions of location probabilities predicted from the estimates in Column 4.

The interactions of the MSA climate and recreation scores with the household location demographics comprise our instruments for the distance between a household's and a stock's headquarters' address. $LogClimate \times LogAge$ and $LogRecreation \times LogAge$ are particularly notable. The respective t-statistics are 4.39 and 4.28. The implied economic effects are that one standard deviation increase in $LogClimate \times LogAge$ increases a household's location probability in a MSA by 9% relative to the average, while one standard deviation increase in $LogRecreation \times LogAge$ increases a household location probability in a MSA by 12% relative to the mean. The above magnitudes make these two interactions be among the strongest predictors of household location choice in our sample.[27]

[27]Of course, there are also other significant interactions for location choice. For example, one of them is $LogHPI \times LogFamSize$, since large families are more likely move to cities with lower house prices. However, since this interaction captures pecuniary location motives, it cannot be excluded for portfolio weights regressions.

4.2. Linear Portfolio Under-Diversification Results

4.2.1. Uncorrected Results

We first estimate the linear under-diversification model of Equation (8). In Table 6, we first show the results of OLS regressions where the endogeneity of *Distance* is not corrected. We run the specification for every month in our sample separately and present the average coefficient estimates along with their respective average t-statistics based on two-way clustered standard errors at the level of the household and the household's MSA.

Regardless of whether we control for stock financial characteristics and household demographics (in Column 1), demographics of the MSAs in which stocks are headquartered (which we add in Column 2), the interactions of the former with the household demographics in the location choice model (e.g. $LogUnemp \times LogAge$) and the (uninteracted) climate and recreation score of the MSAs of the stocks' headquarters (which we incorporate in Column 3), the estimated values of the *Distance* coefficient are very similar. In Column 3, in which we include all controls, the estimated coefficient equals on average -0.101 and has a t-statistic of -5.73.

4.2.2. IV Regressions with One Instrument at a Time

We begin with simple 2SLS regressions, where the endogenous variable of *Distance* is instrumented with only one of the two key excluded location-choice interactions at a time. The estimation results are depicted in Columns 1 and 2 of Table 7. Panels A and B refer to the second and first-stage respectively.

In Column 1, the instrument is the interaction between *LogAge* and *LogClimate*. The just-identified estimate of the *Distance* coefficient is on average equal to -0.082. Thus, there is a 19% decrease relative to the uncorrected case (where the estimated coefficient is -0.101). The average t-statistic is -1.92. Older households locate closer to MSAs with mild climate. Hence, the above interaction which predicts a household's location choice in a MSA

positively (in Column 4 of Table 5), predicts, here, negatively that household's distance from a MSA. In Panel B, the estimated coefficient of the interaction is on average -0.628 and is highly statistically significant, with an average t-statistic of -3.65. The first-stage F statistic is on average 13.45 - i.e., higher than 10 - showing that the instrument is strong (e.g., Stock and Yogo (2002)).[28]

In Column 2, the instrument is the interaction between *LogAge* and *LogRecreation*. The just-identified estimate of the Distance coefficient is on average -0.077 and has an average t-statistic. The reduction relative to the uncorrected case is now 24%, which is a bit higher than before. As we see in Panel B, this instrument is a bit stronger than the previous one. Its average estimated coefficient in the fist-stage is -0.865 and has an average t-statistic of -4.03. The first-stage F-statistic is on average 16.42. The higher magnitudes of the coefficient estimate and the F-statistic point to the 5% additional reduction in the *Distance* coefficient witnessed above.

In Figure 2, we also present the complete time series evolution of the distance coefficient estimates (in percentage points), with and without the correction for location. The OLS estimates are depicted with a blue line, while the IV estimates are depicted with a red line when the instrument is $LogClimate \times LogAge$, and a green line when the instrument is $LogRecreation \times LogAge$. With few exceptions (e.g., first quarter of 1991, second quarter of 1992, etc.), both IV estimates lie quite close to each other.

In Panel C, we also present the results from the OLS reduced form estimation, in which *Distance* is replaced with each one of the excluded household-MSA interactions in the portfolio under-diversification model. In Column 1, the average estimated coefficient of $LogRecreation \times LogAge$ is 0.056, while in Column 2, the average estimated coefficient of $LogRecreation \times LogAge$ is 0.07. The average t-statistics in both Columns are higher than 2, i.e. 2.51 and 2.85 respectively, so that the coefficients of these interactions are statistically significant. The difference between the over-investment in stocks with headquarters in MSAs

[28]The displayed F-statistic is actually the Kleibergen-Paap F-statistic provided by Baum, Schaffer, and Stillman (2007), which accounts for the two-way clustering of the standard errors.

with more mild climate or more recreation opportunities relative to the over-investment in stocks with headquarters in MSAs with less mild climate or fewer recreation opportunities is higher for older investors than younger investors.[29] The reduced form estimation results successfully display the causal relations of interest (Angrist and Krueger (2001)).

4.2.3. IV Regressions With All Instruments

We next present the estimation results from IV regressions that use as instruments all the interactions between the climate and recreation scores in the MSAs of the stocks' headquarters and the household demographics in the location choice model (i.e., age, marital status, number of children, and gender). Recall that household demographics beyond age can also affect the propensity of a household to have a non-pecuniary motive for location - as we discuss above - and hence would be useful to consider as robustness. The 2SLS results are depicted in Column 3 of Table 7 as well as in Column 1 of Table 8. The obtained estimate of the *Distance* coefficient is on average -0.074, which corresponds to a 27% reduction relative to the uncorrected OLS estimate. The average t-statistic is -1.90.

Panel A of Table 7 also contains the p-value of a Hansen J-test for overidentifying restrictions. On average, it equals 0.71, so that the null hypothesis that the instruments are exogenous is not rejected at any reasonable level of statistical significance. In Panel B of the same table, we see that the average value of the first-stage F-statistic increases to 21.37. Our two best instruments, $LogClimate \times LogAge$ and $LogRecreation \times LogAge$, remain strong in the presence of additional excluded interactions. Moreover, in Panel C, we show that the two interactions, in a reduced form in which *Distance* is replaced by all the instruments, have a similar magnitude as in Columns 1 and 2. Their average t-statistics become actually slightly higher.

In Table 8, we compare the average coefficient estimates and t-statistics of 2SLS (in Column 1) with the ones obtained by LIML (in Column 2). Both estimation methods yield

[29] One standard deviation increase in $LogClimate \times LogAge$ ($LogRecreation \times LogAge$) increases a household's excess portfolio weight on a stock by about 9% (15%) relative to the average.

virtually identical results. The coefficient estimate of *Distance* is on average -0.073 and has an average t-statistic of -1.91.

Up to this point, our instrumental variables analysis follows closely the guidelines of Angrist and Pischke (2008). For robustness, we close this section with the GMM estimation of the linear portfolio under-diversification model. The results are tabulated in Column 3 of Table 8. The average coefficient estimate of *Distance* is -0.067 and has an average t-statistic of -1.95. Hence, there is an approximate 34% decrease in the local bias that household portfolios exhibit, i.e. $6\% - 7\%$ more relative to the LIML and 2SLS estimates. The GMM coefficient estimates of local bias over time (in percentage points) are depicted with a yellow line in Figure 2 and lie, almost always, below the red and green lines of the just-identified IV estimates.

4.3. Non-Linear Portfolio Choice Results

4.3.1. Uncorrected Results

We now turn to the non-linear portfolio choice model of Equation (9). Table 9 presents the Tobit estimation results, when location choice is *not* taken into account. In Column 1, we include stock financial characteristics and household demographics as controls. In Column 2, we incorporate all the demographics of the MSAs of the stocks' headquarters, except for climate and recreation. In Column 3, we also interact the latter with the household demographics in the location choice model and additionally control for the scores of climate and recreation of the MSAs in which the stocks are headquartered. Regardless of the observable fundamentals that we introduce in each column, the distance coefficient estimate remains on average intact and equals -0.014. The corresponding value of the t-statistic varies from -5 to -6.[30]

[30]The sign and magnitude of the estimated coefficients of the other controls is as anticipated. Size, book-to-market ratio, turnover and volatility have all positive and statistical significant coefficients. On the other hand, households are not momentum traders, neither do they buy stocks based on their investment, as this is indicated by the negative sign of the respective coefficients. With an almost zero t-statistic, profitability here seems to have no impact on households' portfolio choice. If a household has a high income, investment in a

In every month, based on our estimates, we calculate the marginal effect of *Distance* on the portfolio weight by anchoring all of the covariates at their contemporary mean values. We then define the implied economic effect to be the average of all these monthly marginal effects times a one standard deviation increase in *Distance*. For our model in Column 3 of Table 9 (which features all controls), the economic effect of *Distance* on the portfolio weight is estimated to be -1.96 basis points or -19% of the average household portfolio weight on a Russell 1000 stock (which is 10.31 bps when considering a household's many zero stock positions). Therefore, in line with the Local Bias summary statistics in Table 2, *Distance* is estimated to have a quite sizable economic effect on the household portfolio choice.

4.3.2. Specification of the Control Function

As Equation (16) stands, there is a high dimensionality issue for the model estimation. There are C^2 control functions, $\Psi_{c,h}(\cdot)$, each of which has C probabilities, $\{p_{i,\ell}\}_{\ell=1}^{C}$, as arguments. As a remedy, we implement a robust non-parametric method, in the spirit of Dahl (2002). We adopt the following three identification assumptions:

Assumption 1 (Two Index Sufficiency): The control function has only two arguments, namely the probability with which household i resides in city c and the probability with which it resides in city h:

$$\Psi_{c,h}\left(\{p_{i,\ell}\}_{\ell=1}^{C}\right) = \Psi_{c,h}\left(p_{i,c}, p_{i,h}\right) \tag{17}$$

Assumption 2 (Residence City Independence): The form of the control function does not depend on the residence city c unless $h = c$, i.e.:

$$\Psi_{c,h}\left(p_{i,c}, p_{i,h}\right) = \left\{ \begin{array}{c} \Psi_{c}\left(p_{i,c}\right) \text{ if } h = c \\ \Psi_{h}\left(p_{i,c}, p_{i,h}\right) \text{ if } h \neq c \end{array} \right\} \tag{18}$$

stock becomes more likely and is increased in magnitude. Education and professional industrial proximity also enter positively and statistically significantly.

Assumption 3 (Homogeneity): The form of the control function is not stock headquartered city-specific, i.e.:

$$\begin{aligned} \Psi_c\left(p_{i,c}\right) &= \Psi^s\left(p_{i,c}\right) \\ \Psi_h\left(p_{i,c}, p_{i,h}\right) &= \Psi^d\left(p_{i,c}, p_{i,h}\right) \ \forall h \neq c \end{aligned} \tag{19}$$

According to Assumption 1, only two out of C probabilities are relevant for the impact of location choice on portfolio choice. Namely, the probability that household i locates in the area in which it actually resides, $p_{i,c}$, and the probability that household i locates in the city that has the headquarters of the stock in which it considers investing, $p_{i,h}$. Yet, that assumption still leaves us with C^2 control functions. To this end, we impose Assumption 2, which states that a control function does not depend on the identity of the city in which household i resides. The total number of control functions is then reduced to its square root. Of course, if it happens that stock j is located in the same city as household i does, i.e. $h = c$, then the identity of the residence city becomes relevant again. Lastly, because the total number of cities in our data is large, i.e. $C = 57$, we are still left with a high number of control functions to be estimated from the data. That is why we conveniently impose Assumption 3, which further reduces the control functions to just two. $\Psi^s\left(\cdot\right)$ for the case in which household i considers investing in a stock that is headquartered in the same city in which it resides and $\Psi^d\left(\cdot\right)$ for the case in which the stock's headquarters are located in a different city.[31] Both $\Psi^s\left(\cdot\right)$ and $\Psi^d\left(\cdot\right)$ can be flexibly estimated through a polynomial series expansion.

[31] In short, as in Dahl (2002), Assumptions 1-3 can be thought of as exclusions restrictions on the conditional joint distribution of household i's idiosyncratic investment error, $\epsilon_{i,c,h,j}$, and its maximum order statistic, $v_{i,c}$, given its observed location utilities $\{V_{i,\ell}\}_{\ell=1}^{C}$ (or equivalently, by the monotonicity, the location probabilities $\{p_{i,\ell}\}_{\ell=1}^{C}$), so that:

$$f\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| \{V_{i,\ell}\}_{\ell=1}^{C}\right) = f\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| \{p_{i,\ell}\}_{\ell=1}^{C}\right) = \left\{ \begin{array}{c} f^s\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| p_{i,c}\right) \text{ if } h = c \\ \\ f^d\left(\epsilon_{i,c,h,j}, v_{i,c} \middle| p_{i,c}, p_{ih}\right) \text{ if } h \neq c \end{array} \right\}$$

Then, combining the above equation with Equation (12) yields that:

4.3.3. Corrected Results

In Table 10, we present the Tobit estimation results that correct for location choice. In all columns, we exclude the interactions between the climate and recreation scores of the MSAs of the stocks' headquarters and the household location demographics. These are our instruments and affect household stock-investment only through the location probabilities of the control functions. As in Column 3 of Table 9, the single terms of climate and recreation and the interactions of all *other* demographics of the MSAs of the stocks' headquarters with the household location demographics are included as controls in the Tobit regressions.

In Column 1 of Table 10, we use a cubic approximation for the control functions, while, in Column 2, the approximation is quartic.[32] In any event, the "corrected" coefficient of *Distance* is estimated to be -0.008. This translates into a 43% reduction relative to the estimate in the absence of control functions (which is -0.014).[33] The implied economic effect of distance also drops to -1.06 bps or -10.3% of the average, so that there is an approximate 47% reduction relative to the uncorrected case (where the economic effect is -19% of the average).

Figure 3 depicts the *Distance* coefficient estimates (in percentage points) in the Tobit

$$\mathbb{E}\left(\epsilon_{i,c,h,j}\left|v_{i,c}<0,\left\{V_{i,\ell}\right\}_{\ell=1}^{C}\right.\right)=\left\{\begin{array}{l}\dfrac{\int_{-\infty}^{+\infty}\int_{-\infty}^{0}\epsilon_{i,c,h,j}f^{s}\left(\epsilon_{i,c,h,j},v_{i,c}\left|p_{i,c}\right.\right)d\epsilon_{i,c,h,j}dv_{i,c}}{p_{i,c}}\equiv\Psi^{s}\left(p_{i,c}\right)\text{ if }h=c\\ \dfrac{\int_{-\infty}^{+\infty}\int_{-\infty}^{0}\epsilon_{i,c,h,j}f^{d}\left(\epsilon_{i,c,h,j},v_{i,c}\left|p_{i,c},p_{ih}\right.\right)d\epsilon_{i,c,h,j}dv_{i,c}}{p_{i,c}}\equiv\Psi^{d}\left(p_{i,c},p_{i,h}\right)\text{ if }h\neq c\end{array}\right\}$$

[32] As in Dahl (2002) and Bourguignon, Fournier, and Gurgand (2007), a linear or quadratic approximation of the control functions is not that flexible. In results that are available upon request, we also try a fifth and sixth order polynomial approximation, without finding any further significant decrease in the distance coefficient.

[33] Controlling for location choice in the Tobit model leads to higher t-statistics for the *Distance* coefficient (e.g., from -7 to -20). However, the standard errors of the portfolio choice parameters might not be exact due to the imputation of the location probabilities from the first-stage estimation of the location choice model. This is true in any two-step estimation procedure. Of course, the estimated probabilities are consistent and the sample size in the conditional logit is quite large. Moreover, we run the investment model separately for every month, having different households in every period. The spirit of this exercise resembles bootstrapping.

model over time. The blue line refers to the uncorrected estimates, while the red and green lines refer respectively to the corrected estimates obtained with cubic and quartic polynomial approximations of the correction functions. The graph shows that the above average decrease in the local bias is robust across all periods.

5. Location-Choice Model Residuals and Household Portfolios

Our reduced form analysis points to a natural model of local bias where household portfolio choice depends on (i) expectations of stock-payoffs based on priors which are correlated with the subjective latent expectations in location choice and (ii) private i.i.d. signals whose precision decreases with distance to firm headquarters, in line with the familiarity heuristic. We can can then estimate this model using the residuals from the location choice model as proxies for these priors.

5.1. Model

Specifically, we assume that retail investors are *risk neutral* and have *subjective* beliefs about the one-period ahead returns of stocks. We also assume the existence of a risk-free asset. Investor i has a normal prior for stock j's excess return, $\tilde{f}_j \equiv \tilde{R}_j - R_f$, based on stock j's financial characteristics, $\boldsymbol{X}_j$, and his expectation about stock j's headquarters' city, $L_{i,j}$:

$$\tilde{f}_j \,|\boldsymbol{X}_j \sim \mathcal{N}\left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \nu L_{i,j}, \frac{1}{\tau^0}\right) \tag{20}$$

Investor i receives a private signal for stock j's return:

$$S_{i,j} = \tilde{f}_j + \eta_{i,j} \tag{21}$$

where $\eta_{i,j} | \boldsymbol{X}_j, L_{i,j}, dist_{i,j} \sim \mathcal{N}(0, 1/\tau^{\eta}(dist_{i,j}))$. That is, conditional on stock j's risk factors, $\boldsymbol{X}_j$, investor i's expectation about stock j's headquarters' city, $L_{i,j}$ and the distance between investor i's residence and stock j's headquarters, $dist_{i,j}$, the signal's noise, $\eta_{i,j}$, is assumed to be normally distributed, with zero mean and precision $\tau^{\eta}(dist_{i,j})$. The precision is expected to *decrease* with the distance (i.e., $\tau^{\eta}(dist_{i,j})' < 0$). Consequently, by the projection theorem, investor i's updated subjective expectation for stock j's excess return is:

$$\mathbb{E}\left(\tilde{f}_j | \boldsymbol{X}_j, L_{i,j}, disti, j, S_{i,j}\right) = \alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \nu L_{i,j} + \xi_{i,j} \tag{22}$$

where, since the private signal, $S_{i,j}$, is unobservable to the econometrician, we define investor i's *latent demand* for stock j as:

$$\xi_{i,j} \equiv \frac{\tau^{\eta}(dist_{i,j})}{\tau^0 + \tau^{\eta}(dist_{i,j})} \left[S_{i,j} - (a + \boldsymbol{b}\boldsymbol{X}_j + \nu L_{i,j})\right] \tag{23}$$

The latent demand, $\xi_{i,j}$, reflects investor i's private information about stock j as well as his optimism or pessimism about the stock's prospects. Equations (20) and (21) imply that:

$$\xi_{i,j} | \boldsymbol{X}_j, L_{i,j}, dist_{i,j} \sim \mathcal{N}\left(0, \underbrace{\frac{\tau^{\eta}(dist_{i,j})}{\tau^0\left[\tau^0 + \tau^{\eta}(dist_{i,j})\right]}}_{\sigma^2_{i,j}}\right) \tag{24}$$

so that the variance of investor i's latent demand for stock j (namely, $\sigma^2_{i,j}$) is expected to decrease with the distance between investor i's residence and stock j's headquarters (given that the precision of the signal is conjectured to decrease as well).[34]

By the risk neutrality, Equation (22) is the only relevant moment of stock j's return for investor i's objective, which is to maximize the expected excess return of his wealth under the presence of short-selling constraints and trading costs. Indeed, households do not short, so that investor i's portfolio weight on a stock cannot be negative, i.e. $w_{i,j} \geq 0$ for any $j \in J$.

[34] That is, $\frac{d\sigma^2_{i,j}(dist_{i,j})}{ddist_{i,j}} = \frac{\tau^{\eta}(dist_{i,j})'}{[\tau^0 + \tau^{\eta}(dist_{i,j})]^2} < 0$, if $\tau^{\eta}(dist_{i,j})' < 0$.

Moreover, we assume that investors face quadratic costs from transactions. We specify the quadradicity of transaction costs in terms of acquired market value, so that when investor i acquires $n_{i,j}$ shares of stock j, he pays:

$$TC_{i,j} = \frac{1}{2}\Lambda_i \left(P_j n_{i,j}\right)^2 \tag{25}$$

where Λ_i scales the level of transaction cost that investor i faces for every unit of money spent on stock j. Conveniently, we let that scaling factor be:

$$\Lambda_i = \frac{c}{W_i^2} \tag{26}$$

The more wealth an investor has, the less he has to worry about transaction costs. If the dependence on investor i's wealth, W_i, is quadratic, transaction costs matter much less for richer households and essentially refer to the level of portfolio weights.

Combining Equations (22), (25) and (26), we end up expressing investor i's objective as follows:[35]

$$\max_{\{w_{i,j}\geq 0\}_{j\in\mathcal{J}}} \left\{ \sum_{j\in\mathcal{J}} \left[\left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \nu L_{i,j} + \xi_{i,j}\right) w_{i,j} - \frac{1}{2} c w_{i,j}^2 \right] \right\} \tag{27}$$

The KKT conditions then imply that:

$$w_{i,j} = \left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \nu L_{i,j} + \xi_{i,j}}{c} \right)^+ \tag{28}$$

where $(\cdot)^+ \equiv \max\{\cdot, 0\}$ simultaneously captures investor i's decision of whether and how much to invest in every stock included in his consideration set. According to Equation (28), investor i invests in stock j provided that the expected excess return of the stock is positive in his view. The transaction cost to which he is subject determines the actual level of the

[35] Investor i's budget constraint requires the sum of his stock-portfolio weights to equal one minus the weight on the risk-free asset. The latter is assumed to be perfectly adjustable to the needs of the portfolio optimization problem.

portfolio weight, scaling it up or down.

Since from Equation (24) the latent demand, $\xi_{i,j}$, is normally distributed, Equation (28) constitutes a heteroskedastic Tobit model. However, as the model stands, it is under-identified. To this end, we make the following simplifications. First, we normalize investors' precision, τ^0 and transaction cost parameter, c, to be equal to 1. And second, we parameterize the precision of the signal's noise, $\eta_{i,j}$, as follows:

$$\tau^{\eta}(dist_{i,j}) = \frac{\exp[2\left(\gamma + \delta dist_{i,j}\right)]}{1 - \exp[2\left(\gamma + \delta dist_{i,j}\right)]} \tag{29}$$

Equation (24) together with (28) entails that:

$$w_{i,j} = \left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \nu L_{i,j} + \exp\left(\gamma + \delta dist_{i,j}\right)\zeta_{i,j}\right)^{+} \tag{30}$$

where $\zeta_{i,j}\,|\boldsymbol{X}_j, L_{i,j}, dist_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Equation (30) is the standard textbook expression for a Tobit model with heteroskedasticity (e.g., Wooldridge (2010)). Here, the heteroskedasticity is specified in terms of the distance of investor i's residence from stock j's headquarters' city.

As a proxy for investor i's priors about stock j's headquarters' city, we use the Pearson residuals from the conditional logit model in Section 2. That is, we define:

$$L_{i,j} \equiv \frac{r_{i,j} - p_{i,j}}{\sqrt{p_{i,j}}} \tag{31}$$

where $r_{i,j}$ equals 1 if investor i resides in stock j's headquarters' city and 0 otherwise, while $p_{i,j}$ is the investor i's predicted location probability in that city.

5.2. Estimation

We estimate the model for every month in our sample separately. For comparison, we also estimate two restricted versions that the model nests (henceforth referred to as Full Model). In the first version, the coefficient of investor i's priors about stock j's headquarters' city,

$L_{i,j}$, equals zero (henceforth referred to as No LocRes - because there is no location residual). In the second version, the coefficient of the distance between investor i's residence and stock j's headquarters, $dist_{i,j}$, equals zero (henceforth referred to as No Distance). We present the time series averages of the coefficients in Panel A of Table 11, along with the average t-statistics based on two-way clustered standard errors at the level of the investor and the investor's MSA.

As reported in Column 1 of Panel A, the estimated coefficient of the location residual (LocRes) is on average 0.126, with an average t-statistic of 16.75. On the other hand, the estimated coefficient of Distance is on average equal to -0.002, with an average t-statistic of approximately -3.04. In Column 2, where the coefficient of LocRes is restricted to be zero, the average estimated coefficient of Distance is -0.004 (i.e., 50% times larger in magnitude than in the Full Model), with an average t-statistic of -5.65. In Column 3, where the coefficient of Distance is constrained at zero, LocRes has an average estimated coefficient equal to 0.138 (i.e., about 10% higher than in the Full Model) and an average t-statistic of 15.55.[36]

Panel B of Table 11 depicts the economic effects of LocRes and Distance. In Column 1 (which corresponds to the estimates in the Full Model), the estimated economic effect of LocRes is on average 1.5 bps or 14.5% of the mean portfolio weight on a stock in Russell 1000. In the same column, the economic effect of Distance is on average -0.98 bps or -9.5% of the mean. Hence, the economic effect of Residual (which expresses the latent expectations from the locational decisions) is bigger than the economic effect of Distance. Household priors account for 60% of local bias and familiarity or distance 40%.

In Column 2 (which refers to the estimates in the No LocRes), the economic effect of Distance equals on average -2.25 bps or -21.8% of the mean. That is, its magnitude is about 130% larger than in the Full Model. In Column 3 (which corresponds to the results in

[36]The estimated coefficients of the stock characteristics are very similar across the three different models. In results that are available upon request, we also experiment with specifications in which we include as additional controls household demographics as well as demographics of the MSAs of the stocks' headquarters. The coefficients of LocRes and Distance that we obtain are very similar.

the No Distance), the economic effect of Residual is on average 1.68 bps or 12% of the mean. This figure is about 11% higher than the economic effect of Residual in the Full Model.

5.3. Model Fit

According to the estimation results of the Full Model, the coefficients of LocRes and Distance are statistically significant. In this subsection, we also perform a nested likelihood ratio tests for hypotheses according to which only one of the two coefficients is statistically significant.

We test the null hypothesis $H_0 : \nu = 0$ (No LocRes) against the alternative hypothesis $H_1 : \nu \neq 0$ (Full Model). Under the null, investor i's priors about stock j's headquarters' city, $L_{i,j}$, does not affect his portfolio weight on the stock. The test is performed by calculating the test statistic $D = 2[\log(L_{H_1}) - \log(L_{H_0})]$ (i.e., twice the difference between the log-likelihood in the Full Model and the log-likelihood in No LocRes) in every month. We then average over time the monthly test statistics to obtain the value of the D-statistic in respective sample period. The calculated value is remarkably high (i.e. $2,185.5$). To test H_0, we compare this figure to critical values of the χ^2 distribution with 1 degree of freedom (e.g., 10.83 at the $\alpha = 0.1\%$ level of statistical significance). We reject the null hypothesis at any reasonable value of statistical significance, confirming that investors' latent expectations about the cities in which they can locate affect their subsequent investment decisions.

In the same spirit, we test the null hypothesis $H_0 : \delta = 0$ (No Distance) against the alternative $H_1 : \delta \neq 0$ (Full Model). According to this null, the distance investor i's residence and stock j's headquarters' city, $dist_{i,j}$, does not affect investor i's precision regarding stock j's signal. The value of the D-statistic is now 212.48. This figure is again much higher than critical values of the χ^2 distribution with 1 degree of freedom, at any reasonable level of statistical significance. Hence, H_0 is rejected, showing that geographical proximity has a statistically significant effect on the investor's portfolio choices.

6. Conclusion

This local bias puzzle is generally explained by theories that assign a causal role to proximity. The empirical analyses typically assume that households locate randomly. But a household in practice optimally locates in a city depending latent subjective expectations about the economic prospects of a city, which correlated with demand for local stocks. We propose a correction for this selection bias in reduced-form portfolio regressions using location choice models. We then propose and estimate a structural model of local bias driven by household priors about stocks in a city, using location-choice model residuals, and familiarity, whereby confidence about stock-payoff signals rise with proximity.

Our analysis points to several future research paths. First, the household finance literature has focused on how observable household (e.g., education) or asset characteristics (e.g., proximity to household) might influence or bias portfolio decisions. Less explored are latent expectations embedded in locational choice decisions. While we have focused on stocks, our analysis naturally applies to general portfolio construction including purchases of homes. Second, we have shown how latent expectations regarding location choice are important for understanding local bias. Naturally, local bias at the MSA level can contribute to international home equity bias. It would be interesting to understand the extent to which such latent expectations also play a role for this other well-known puzzle.

Appendix: Definition of the MSA Livability Scores

Transportation: This score is calculated based on the daily commute, public transportation, national highways, air service and passenger rail service. The higher the score of transportation, the better the transportation in the MSA.

Colleges: This score is based on the number of students enrolled in community or two-year colleges, the number of students enrolled in private four-year and graduate-level institutions and the number of students enrolled in public four-year and graduate level institutions. The higher the score of colleges, the better the colleges in the MSA.

Health Care: This score is based on the number of general/family practitioners per 100K population, the number of medical specialists per 100K population, the number of surgical specialists per 100K population and the number of hospitals approved for physician residency programs by the AMA. The higher the score of health care, the better the health care in the MSA.

Crime: This score is based on the violent crime rate and the property crime rate divided by 10. The lower the score of crime, the less the crime in the MSA.

Recreation: This score is based on the number of public golf courses, good restaurants, movie theater screens, zoos, aquariums, family theme parks, parimutuel betting attractions, professional sports, collegiate sports, miles of ocean or Great Lakes coastline, national forests, national parks and national wildlife refuges and state or provincial parks. The higher the score of recreation, the better the recreation in the MSA.

Climate: The score is based on the number of very hot and cold months, the seasonal temperature variation, the number of heating and cooling degree days, the number of freezing days, the number zero-degree days and the number of 90-degree days. The higher the score of climate, the better the climate in the MSA.

References

Agarwal, S., J. C. Driscoll, X. Gabaix, and D. Laibson, 2009, "The Age of Reason: Financial Decisions over the Life Cycle and Implications for Regulation," *Brookings Papers on Economic Activity*, 2009(2), 51–117.

Angrist, J. D., and A. B. Krueger, 2001, "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 15(4), 69–85.

Angrist, J. D., and J.-S. Pischke, 2008, *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton University Press.

Bajari, P., and M. E. Kahn, 2005, "Estimating Housing Demand With an Application to Explaining Racial Segregation in Cities," *Journal of Business and Economic Statistics*, 23(1), 20–33.

Barber, B. M., and T. Odean, 2000, "Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors," *Journal of Finance*, 55(2), 773–806.

Baum, C. F., M. E. Schaffer, and S. Stillman, 2007, "Enhanced Routines for Instrumental Variables/GMM Estimation and Testing," *Stata Journal*, 7(4), 465–506.

Bayer, P., F. Ferreira, and R. McMillan, 2007, "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy*, 115(4), 588–638.

Bayer, P. J., B. D. Bernheim, and J. K. Scholz, 2009, "The Effects of Financial Education in the Workplace: Evidence from a Survey of Employers," *Economic Inquiry*, 47(4), 605–624.

Bishop, K. C., 2007, "A Dynamic Model of Location Choice and Hedonic Valuation," *Working Paper*.

Bourguignon, F., M. Fournier, and M. Gurgand, 2007, "Selection Bias Corrections Based on the Multinomial Logit Model: Monte Carlo Comparisons," *Journal of Economic Surveys*, 21(1), 174–205.

Brandt, W. M., P. Santa-Clara, and R. Valkanov, 2009, "Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns," *Review of Financial Studies*, 22(9), 3411–3447.

Campbell, J. Y., 2006, "Household Finance," *The Journal of Finance*, 61(4), 1553–1604.

Charles, K. K., E. Hurst, and N. Roussanov, 2009, "Conspicuous Consumption and Race," *The Quarterly Journal of Economics*, 124(2), 425–467.

Clark, D. E., and W. J. Hunter, 1992, "The Impact of Economic Opportunity, Amenities and Fiscal Factors on Age-Specific Migration Rates," *Journal of Regional Science*, 32(3), 349–365.

Coval, J. D., and T. J. Moskowitz, 1999, "Home Bias at Home: Local Equity Preference in Domestic Portfolios," *Journal of Finance*, 54(6), 2045–2073.

——— , 2001, "The Geography of Investment: Informed Trading and Asset Prices," *Journal of Political Economy*, 109(4), 811–841.

Dahl, G. B., 2002, "Imobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica*, 70(6), 2367–2420.

DeMarzo, P. M., R. Kaniel, and I. Kremer, 2004, "Diversification as a Public Good: Community Effects in Portfolio Choice," *Journal of Finance*, 59(4), 1677–1716.

Diamond, R., 2016, "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000," *American Economic Review*, 106(3), 479–524.

Dubé, J.-P., X. Luo, and Z. Fang, 2017, "Self-Signaling and Prosocial Behavior: A Cause Marketing Experiment," *Marketing Science*, 36(2), 161–186.

Fama, E. F., and K. R. French, 1992, "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47(2), 427–465.

Feng, L., and M. S. Seasholes, 2008, "Individual Investors and Gender Similarities in An Emerging Stock Market," *Pacific-Basin Finance Journal*, 16, 44–60.

French, K. R., and J. M. Poterba, 1991, "Investor Diversification and International Equity Markets," *American Economic Review*, 81(2), 222–226.

Gargano, A., and A. G. Rossi, 2018, "Does it Pay to Pay Attention?," *Review of Financial Studies*.

Goetzmann, W. N., and A. Kumar, 2008, "Equity Portfolio Diversification," *Review of Finance*, 12(3), 433–463.

Gomez, J.-P., R. Priestley, and F. Zapatero, 2009, "Implications of Keeping-Up-with-the-Joneses Behavior for the Equilibrium Cross Section of Stock Returns: International Evidence," *Journal of Finance*, 64(6), 2703–2737.

Grinblatt, M., and M. Keloharju, 2001, "How Distance, Language, and Culture Influence Stockholdings and Trades," *Journal of Finance*, 56(3), 1053–1073.

Heath, C., and A. Tversky, 1991, "Preference and Belief: Ambiguity and Competence in Choice under Uncertainty," *Journal of risk and uncertainty*, 4(1), 5–28.

Heckman, J. J., 1977, "Sample Selection Bias as a Specification Error (with an Application to the Estimation of Labor Supply Functions)," *NBER Working Papers*.

Hizmo, A., 2015, "Risk in Housing Markets: An Equilibrium Approach," *Working Paper*.

Hong, H., W. Jiang, N. Wang, and B. Zhao, 2014, "Trading for Status," *Review of Financial Studies*, 27, 3171–3212.

Huberman, G., 2001, "Familiarity Breeds Investment," *Review of Financial Studies*, 14(3), 659–680.

Ivković, Z., and S. Weisbenner, 2005, "Local Does as Local Is: Information Content of the Geography of Individual Investors' Common Stock Investments," *The Journal of Finance*, 60(1), 267–306.

Kaplan, G., and S. Schulhofer-Wohl, 2012, "Understanding the Long-Run Decline in Interstate Migration," working paper, National Bureau of Economic Research.

Keloharju, M., S. Knupfer, and E. Rantapuska, 2012, "Mutual Fund and Share Ownership in Finland," *Liiketaloudellinen aikakauskirja*, 2, 178–198.

Kennan, J., and J. R. Walker, 2011, "The Effect of Expected Income on Individual Migration Decisions," *Econometrica*, 79(1), 211–251.

Korniotis, G. M., and A. Kumar, 2011, "Do Older Investors Make Better Investment Decisions?," *Review of Economics and Statistics*, 93(1), 244–265.

Lusardi, A., and O. Mitchell, 2007, "Financial Literacy and Retirement Preparedness: Evidence and Implications for Financial Education," *Business Economics*, 42(1), 35–44.

Luttmer, E. F. P., 2005, "Neighbors as Negatives: Relative Earnings and Well-Being," *The Quarterly Journal of Economics*, 120(3), 963–1002.

Massa, M., and A. Simonov, 2006, "Hedging, Familiarity and Portfolio choice," *Review of Financial Studies*, 19(2), 633–685.

Novy-Marx, R., 2013, "The Other Side of Value: The Gross Profitability Premium," *Journal of Financial Economics*, 108(1), 1–28.

Odean, T., 1999, "Do Investors Trade Too Much?," *American Economic Review*, 89(5), 1279–1298.

Ortalo-Magné, F., and A. Prat, 2016, "Spatial Asset Pricing: A First Step," *Economica*, 83, 130–171.

Petersen, M. A., 2009, "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," *Review of Financial Studies*, 22(1), 435–480.

Pirinsky, C., and Q. Wang, 2006, "Does Corporate Headquarters Location Matter for Stock Returns?," *Journal of Finance*, 61(4), 1991–2015.

Roberts, M. R., and T. M. Whited, 2013, "Endogeneity in Empirical Corporate Finance," in *Handbook of the Economics of Finance*. Elsevier, vol. 2, pp. 493–572.

Savageau, D., and R. Boyer, 1993, *Places Rated Almanac: Your Guide to Finding the Best Places to Live in North America*. Prentice Hall.

Seasholes, M. S., and N. Zhu, 2010, "Individual Investors and Local Bias," *The Journal of Finance*, 65(5), 1987–2010.

Sharpe, W. F., 1964, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, 19(3), 425–442.

Stock, J. H., and M. Yogo, 2002, "Testing for Weak Instruments in Linear IV Regression," .

Tuzel, S., and M. B. Zhang, 2017, "Local Risk, Local Factors, and Asset Prices," *Journal of Finance*, 72(1), 325–370.

Wooldridge, J. M., 2010, "Econometric Analysis of Cross Section and Panel Data," Cambridge, MA: MIT Press.

Zhu, N., 2002, "The Local Bias of Individual Investors," *Yale ICF working paper*.

Table 1: Summary Statistics

This table reports the summary statistics of all the variables in our sample (in levels and in logs). Panel A refers to the MSA demographics. IncPerCap is the income per capita. Unemp is the unemployment rate. HPI is the housing price index. Pop is the population number. Transportation is the score for the ability to meet transportation needs. Colleges is the score for the college opportunities. Healthcare is the score for the supply of health care. Crime is the score for the crime. Recreation is the score for the supply of recreation assets. Climate is the score for the climate mildness. Panel B refers to the household demographics. Income is the income of the household's head. Education is the percentage of the household's ZIP-Code population that holds a bachelor's or higher degree. ProfProxim is the professional industrial proximity to a stock of the household's ZIP-Code population. Professional, Managerial, SalesSvc, WhiteCollar and BlueCollar refer to the occupation code of the household's head and are indicator variables that equal one if the household's head has a professional, managerial, sales-services, white collar or blue collar-type job. White, Black, Hispanic, AsianOth are respectively the percentages of the household's ZIP-Code population that are white, black, Hispanic, Asian or of other race. The list of household demographics for location choice consists of: Age, i.e. the age of the household head, FamSize, i.e. the family size of the household, Male and Married, which are respectively indicator variables that equal one if the household's head is a male or married. Panel C refers to the household stock holdings. Portval is the portfolio value of a household. Numstk is the number of stocks that a household holds. Portwt is the portfolio weight of a household on a stock at the extensive margin. EW is the excess household portfolio weight (relative to the market). Panel D refers to the stock geographical proximity in the investment universe. Distance is the distance (in degrees) between a household's address ZIP-Code and the address ZIP-Code of a stock's headquarters. Away250m (Away100m) is an indicator variable that equals one if a stock is headquartered more than 250 miles (100 miles) away from a household's residence. Panel E refers to the financial characteristics of the Russell 1000 stocks. Size is the market capitalization. BTM is the book-to-market ratio. Turnover is the monthly share turnover. Momentum is the past 12-month return. Volatility is the volatility of the monthly returns in the past 12 months. Profitability is the ratio of past annual gross profits to assets. Investment is the past annual growth rate of assets. Food to OthSvc are indicator variables that equal one if a stock belongs to the corresponding industry. The industries are defined based on the 17 Fama-French industry portfolios. The sample period is from January 1991 to November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

Panel A: MSA Demographics

	Mean	S.D.	Median	Min	Max
IncPerCap (thousand $)	21.63	3.31	20.84	16.96	44.88
LogIncPerCap	3.06	0.14	3.04	2.83	3.8
Unemp (%)	6.25	1.71	5.97	2.92	15.63
LogUnemp	1.8	0.25	1.79	1.07	2.75
HPI	94.8	10.63	94.32	66.04	122.63
LogHPI	4.55	0.11	4.55	4.19	4.81
Pop (million)	2.46	2.71	1.52	0.68	17.68
LogPop	14.4	0.71	14.23	13.43	16.69
Transportation	4,816.47	1,235.59	4,705	6.97	7,429
LogTransport	8.36	0.89	8.46	1.94	8.91
Colleges	1,523.38	1,400.53	1,122	1.45	6,728
LogColleges	6.88	1.26	7.02	0.37	8.81
HealthCare	298.14	140.27	274	109	880
LogHealthCare	5.61	0.39	5.61	4.69	6.78
Crime	1,219.02	399.44	1,165	550	2,821
LogCrime	7.06	0.31	7.06	6.31	7.94

Table Cont'd: Summary Statistics

Panel A Cont'd: MSA Demographics					
	Mean	S.D.	Median	Min	Max
Recreation	2,130.49	787.78	2,104	707	3,940
LogRecreation	7.59	0.39	7.65	6.56	8.28
Climate	577.02	116.71	559	287	910
LogClimate	6.34	0.2	6.33	5.66	6.81
Panel B: Household Demographics					
	Mean	S.D.	Median	Min	Max
Income (thousand $)	101.85	68.24	87.5	10	250
LogIncome	11.33	0.65	11.38	9.21	12.43
Education	0.36	0.15	0.35	0	0.91
ProfProxim	0.08	0.06	0.07	0	0.56
HH Occupation-Codes					
Professional	0.56	0.5	1	0	1
Managerial	0.27	0.44	0	0	1
SalesSvc	0.08	0.28	0	0	1
WhiteCollar	0.05	0.22	0	0	1
BlueCollar	0.04	0.19	0	0	1
HH ZIP-Code Race Pct.					
White	0.81	0.17	0.87	0	1
Black	0.06	0.1	0.02	0	0.98
Hispanic	0.07	0.09	0.04	0	0.94
AsianOth	0.06	0.08	0.04	0	0.62
HH Location Demo's					
Age	51.78	12.51	50	21	80
LogAge	3.92	0.24	3.91	3.04	4.38
FamSize	2.49	1.2	2	1	6
LogFamSize	0.8	0.48	0.69	0	1.79
Male	0.92	0.28	1	0	1
Married	0.73	0.44	1	0	1
Panel C: Household Stock Holdings					
	Mean	S.D.	Median	Min	Max
Portval ($)	30,776.79	126,247.43	11,255.43	1,000	16,227,021
Numstk	2.32	2.27	1.7	1	36
portwt	10.31 bps	0.03	0	0	1
EW	1.07	161.52	-1	-1	1,648,400
Panel D: Stock Geographical Proximity					
	Mean	S.D.	Median	Min	Max
Distance (degrees)	17.6	11.73	15.48	0	39.72
Away250	0.87	0.33	1	0	1
Away100	0.93	0.26	1	0	1

Table Cont'd: Summary Statistics

Panel D: Russell 1000 Stock Characteristics					
	Mean	S.D.	Median	Min	Max
Size (million $)	3589.48	7962.3	1245.65	1.7	159758.9
LogSize	7.25	1.29	7.13	0.53	11.98
BTM	0.56	0.81	0.49	-31.45	29.06
Turnover	0.1	0.12	0.06	0	1.86
Momentum	0.12	0.52	0.06	-0.97	16.74
Volatility	0.09	0.05	0.08	0.01	1.76
Profitability	0.34	0.28	0.27	-0.58	2.1
Investment	0.2	0.72	0.08	-0.94	33.94
Food	0.04	0.19	0	0	1
Mines	0.02	0.13	0	0	1
Oil	0.05	0.22	0	0	1
Clths	0.02	0.12	0	0	1
Durbl	0.02	0.13	0	0	1
Chems	0.03	0.17	0	0	1
Cnsum	0.04	0.21	0	0	1
Cnstr	0.02	0.14	0	0	1
Steel	0.02	0.14	0	0	1
FabPr	0.01	0.1	0	0	1
Machn	0.10	0.31	0	0	1
Cars	0.01	0.12	0	0	1
Trans	0.04	0.19	0	0	1
Utils	0.07	0.26	0	0	1
Rtail	0.07	0.26	0	0	1
Finan	0.18	0.39	0	0	1
OthSvc	0.25	0.44	0	0	1

Table 2: Local Bias among Households

This table provides summary statistics for the local bias of household portfolio holdings in our data. Column 2 (Avg. Distance from Holdings) reports the average distance of households from the stocks they hold in their portfolios. The average distance in Column 2 is computed as $\frac{1}{I}\sum_i \sum_j w_j^i d_j^i$, where d_j^i is the distance between household i's residential area and the headquarters area of stock j, w_j^i is household i's portfolio weight on stock j, and I is the total number of households. Column 3 (Avg. Distance from Benchmark) is computed as $\frac{1}{I}\sum_i \sum_j \bar{w}_j d_j^i$, where $\bar{w}_j$ is a benchmark Russell 1000 portfolio weight on stock j, and d_j^i and I are the same as in Column 2. In Row 1 the benchmark is the equally weighted portfolio. In Row 2 the benchmark is the value-weighted portfolio. Column 4 (Difference) reports the difference between Column 2 and Column 3, which is the local bias in distance units. Column 5 (% Bias (LB)) reports the local bias (LB) measure as a percentage. Column 6 (t-stat) reports the t-statistics for the LB measure. The sample period is from January 1991 to November 1996.

	Avg. Distance from				
	(1)	(2)	(3)	(4)	(5)
Weights	Holdings	Benchmark	Difference	% Bias (LB)	t-stat
Equal	9.38	17.67	8.29	45.45	52.74
Value	9.38	17.65	8.26	43.72	47.65

Table 3: MSA Demographics Balance Test Based on Climate and Recreation

This table presents the MSA demographics balance test based on recreation and climate. In Panel A, the MSA sample is split into two groups based on the median climate score. In Panel B, the MSA sample is split into two groups based on the median recreation score. In each subsample, we calculate the averages of the MSA demographics. Column 1 refers to the subsamples in which the score of recreation (in Panel A) or climate (in Panel B) are below the median. Column 2 refers to the subsamples above the corresponding median. Column 3 depicts the differences between the average MSA demographics in the two groups. Column 4 depicts the t-statistics of paired difference tests. Since annual data (from 1991-1996) are available for the MSAs' income per capita, HPI, unemployment rate and population, we run the balance test for these variables in every year and present the time-series average means, differences and t-statistics. The metropolitan area sample consists of 57 MSAs with a population of at least 750K at the end of 1996.

	Panel A: Split of MSAs based on Climate			
Averages	(1) Below Median	(2) Above Median	(3) Difference	(4) t-statistic
IncPerCap (thousand $)	22.62	24.66	2.04	1.82
HPI	97.92	100.5	2.58	0.9
Unemp (%)	5.58	6.02	0.44	0.91
Pop (million)	1.98	3.09	1.11	1.52
Transportation	4,605.45	5,082.96	477.51	1.51
Colleges	1,350.07	1,798.96	448.89	1.24
Crime	1,237.86	1,205.64	-32.22	-0.3
Healthcare	265.52	328.46	62.94	1.4
Recreation	2,133.34	2,269.29	135.95	0.67

	Panel B: Split of MSAs based on Recreation			
Averages	(1) Below Median	(2) Above Median	(3) Difference	(4) t-statistic
IncPerCap (thousand $)	23.94	23.29	-0.65	-0.68
HPI	99.2	99.17	-0.03	-0.06
Unemp (%)	5.43	6.17	0.74	1.58
Pop (million)	2.29	2.77	0.48	0.65
Transportation	4,992.41	4,682.18	-310.23	-0.97
Colleges	1,406.28	1,740.75	334.47	0.92
Crime	1,158.83	1,287.5	128.67	1.22
Healthcare	289.69	303.43	13.74	0.46
Climate	572.34	592.54	20.2	0.64

Table 4: Stock Financial Characteristics Balance Test Based on Climate and Recreation in the MSAs of the Headquarters

This table presents the estimation results of two-variable regressions that test the "balance" of stock financial characteristics with respect to the recreation and climate score in the MSAs where the stocks are headquartered. The estimation is performed in a cross-section of stocks for every month separately. The depicted results are the average monthly coefficient estimates (in Columns 1 and 3) and the average t-statistics based on clustered standard errors at the level of the MSA of the stocks' headquarters (in Columns 2 and 4). The listed stock financial characteristic is the dependent variable in the regressions. In Columns 1 and 2, the independent variable is LogClimate. In Columns 3 and 4, the independent variable is LogRecreation.

	Independ. Variable: LogClimate		Independ. Variable: LogRecreation	
	(1)	(2)	(3)	(4)
Depend. Variable	Coef. Est.	t-statistic	Coef. Est.	t-statistic
LogSize	-0.065	-0.28	-0.142	-1.31
BTM	-0.001	-0.26	0.039	0.92
Turnover	0.098	1.58	0.002	0.18
Momentum	0.086	0.78	0.009	0.21
Volatility	0.029	1.5	0.009	[0.99]
Profitability	0.089	0.86	-0.009	-0.22
Investment	0.131	1.29	0.037	0.3
Food	-0.078	-1.8	-0.002	-0.1
Mines	-0.005	-0.29	0.006	0.64
Oil	-0.136	-1.13	0.038	0.73
Clths	0.015	1.15	-0.001	-0.11
Durbl	-0.026	-1.68	-0.01	-0.77
Chems	-0.003	-0.22	-0.034	-1.59
Cnsum	0.046	1.38	0.01	0.6
Cnstr	-0.001	-0.07	-0.024	-1.47
Steel	0.008	0.48	-0.012	-0.84
FabPr	-0.015	-1.63	0.001	0.22
Machn	0.182	1.23	-0.07	-0.89
Cars	-0.001	-0.07	0.01	1.42
Trans	-0.001	-0.03	0.002	0.1
Utils	-0.069	-1.45	0.006	0.23
Rtail	-0.032	-0.8	0.006	0.21
Finan	0.006	0.08	0.016	0.31
OthSvc	0.11	1.7	0.058	1.62

Table 5: Conditional Logistic Regression of Household Location Choice

This table presents the results from the maximum likelihood estimation of four conditional logit models of household location choice. The table shows the coefficient estimates and t-statistics [in brackets] based on standard errors clustered at the household level. The choice set of households consists of 57 MSAs with a population of at least 750K. The dependent variable is a household-city-specific indicator variable that equals one if the household resides in the specific MSA. In Columns 1 and 3, the explanatory variables consist of the following MSA demographics in logs: the income per capita (LogIncPerCap), the unemployment rate (LogUnemp), the house price index (LogHPI), the population number (LogPop) and the scores of transportation (LogTransport), colleges (LogColleges), crime (LogCrime) and health care (LogHealthCare). In Columns 2 and 4, the explanatory variables include also logs of the scores of the MSA climate mildness (LogClimate), and recreation (LogRecreation). In Columns 3 and 4, pair-wise interaction terms based on the available MSA demographics and household location demographics are added as controls. The list of household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable that equals one if the head is male. For brevity, only the interactions of the MSA log scores of climate and recreation with the log of the household head's age are being depicted in Column 4. Log-likelihood is the log of the likelihood of the conditional logit model. A likelihood ratio test is performed in Columns 1, 2 and 3 for restricted versions of the full model in Column 4. Pseudo R^2 is McFadden's pseudo "R-square" measure based on the estimated log-likelihood value. AIC is the Akaike information critetion.

	Homogeneity		Matching	
	(1)	(2)	(3)	(4)
LogIncPerCap	2.984	3.083	14.865	13.007
	[24.58]	[23.57]	[6.9]	[5.57]
LogUnemp	-0.559	-0.445	0.146	-0.135
	[-8.97]	[-7.29]	[0.13]	[-0.12]
LogHPI	0.089	-0.969	-5.843	-7.682
	[0.68]	[-6.95]	[-2.5]	[-3.09]
LogPop	0.878	0.834	0.773	1.219
	[16.08]	[14.92]	[0.78]	[1.2]
LogTransport	-0.067	-0.016	-0.052	-0.122
	[-3.21]	[-0.77]	[-0.16]	[-0.37]
LogColleges	0.201	0.227	1.061	1.003
	[4.88]	[5.45]	[1.45]	[1.36]
LogCrime	0.163	0.095	-1.151	-1.193
	[3.24]	[1.86]	[-1.26]	[-1.29]
LogHealthCare	-0.486	-0.555	-4.148	-4.522
	[-7.46]	[-8.08]	[-3.55]	[-3.69]
LogClimate		0.783		3.243
		[11.42]		[2.59]
LogRecreation		0.27		-1.772
		[7]		[-2.6]
× HH Location Demo's	NO	NO	YES	YES
LogClimate × LogAge				0.436
				[4.39]
LogRecreation × LogAge				0.571
				[4.28]
Log-Likelihood	-31,128	-30,956	-31,032	-30,835
Pseudo R^2	0.113	0.119	0.116	0.122
LR-Test vs. Full Model	583.12	240.86	392.76	-
AIC	62,270	61,931	62,143	61,770

Table 6: OLS Regressions of Household Excess Portfolio Weights, No Correction for Location Choice

This table presents the estimation results from three OLS regressions of household excess portfolio weights. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $EW_{i,j} \equiv \frac{w_{i,j}-w_j^{VW}}{w_j^{VW}}$, i.e. the excess portfolio weight of household i on stock j (relative to the market value-weighted portfolio weight on stock j). The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP-Code from household i's address ZIP-Code. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. In Column 1, the controls consist of the stock financial characteristics and the household demographics. In Column 2, the demographics of the MSAs of the stocks' headquarters are added as controls. In Column 3, the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are also added as controls. The list of the household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The constant term in the OLS regressions are not reported for brevity. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

	Dependent Variable: Excess Portfolio Weight $EW_{i,j} \equiv \frac{w_{i,j}-w_j^{VW}}{w_j^{VW}}$		
	(1)	(2)	(3)
Distance	-0.099	-0.101	-0.101
	[-5.1]	[-5.72]	[-5.73]
LogSize	-0.595	-0.608	-0.608
	[-5.53]	[-5.82]	[-5.82]
BTM	-0.35	-0.368	-0.367
	[-0.25]	[-0.32]	[-0.32]
Turnover	3.328	2.669	2.669
	[2.73]	[2.21]	[2.21]
Momentum	-1.541	-1.557	-1.557
	[-5.65]	[-5.64]	[-5.64]
Volatility	30.972	29.864	29.863
	[5.57]	[5.38]	[5.38]
Profitability	-0.297	-0.495	-0.494
	[-0.4]	[-0.94]	[-0.93]
Investment	-0.51	-0.526	-0.526
	[-3.31]	[-3.34]	[-3.34]
LogIncome	-0.04	-0.038	-0.038
	[-0.31]	[-0.3]	[-0.3]
Education	-0.634	-0.631	-0.633
	[-1.24]	[-1.22]	[-1.23]
ProfProxim	8.108	8.127	8.071
	[4.17]	[4.23]	[4.25]
Stock Industry FE	YES	YES	YES
HH Occupation-Code FE	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES
HH Location Demo's	YES	YES	YES
Stock MSA Demo's	NO	YES	YES
× HH Location Demo's	NO	NO	YES
Stock MSA Climate	NO	NO	YES
Stock MSA Recreation	NO	NO	YES

Table 7: 2SLS and Reduced Form Regressions of Household Excess Portfolio Weight on Distance

This table presents the estimation results from three two-stage least squares regressions and three reduced-form regressions of household excess portfolio weights. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $EW_{i,j} \equiv \frac{w_{i,j}-w_j^{VW}}{w_j^{VW}}$, i.e. the excess portfolio weight of household i on stock j (relative to the market value-weighted portfolio weight on stock j). The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP Code from household i's address. Household demographics, stock financial characteristics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. Panel A (B) presents the second-stage (first-stage) estimation results. Panel C presents the reduced form estimation results (when distance is replaced with the instruments). The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. In Columns 1 and 2, only a single instrument is used. *LogClimate* × *LogAge* in Column 1 and *LogRecreation* × *LogAge* in Column 2. In Column 3, all instruments are used. The F-statistic tests the hypothesis that the coefficient(s) of the instrument(s) are (jointly) zero - taking into account the two-level clustering. The Hansen J-test is the test of overidentifying restrictions. The (log) climate and recreation scores of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. The list of household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

	Using One Instrument		ALL
	(1)	(2)	(3)
Panel A (*Second-Stage*): Dependent Variable: Excess Portfolio Weight $EW_{i,j} \equiv \frac{w_{i,j}-w_j^{VW}}{w_j^{VW}}$			
Distance	-0.082	-0.077	-0.074
	[-1.92]	[-1.84]	[-1.90]
Controls (for *all* panels)			
Stock Financial Char's	YES	YES	YES
Household Demo's	YES	YES	YES
Stock MSA Demo's	YES	YES	YES
× HH Location Demo's	YES	YES	YES
Stock MSA Climate	YES	YES	YES
Stock MSA Recreation	YES	YES	YES
Hansen J-test (p-value)			0.71
Panel B (*First-Stage*): Dependent Variable: Distance $dist_{i,j}$			
LogClimate × LogAge	-0.628		-0.516
	[-3.65]		[-3.99]
LogRecreation × LogAge		-0.865	-0.75
		[-4.03]	[-3.82]
Other Instruments	NO	NO	YES
F-statistic	13.45	16.42	21.37
Panel C (*Reduced Form*): Dependent Variable: Excess Portfolio Weight $EW_{i,j} \equiv \frac{w_{i,j}-w_j^{VW}}{w_j^{VW}}$			
LogClimate × LogAge	0.056		0.047
	[2.51]		[2.88]
LogRecreation × LogAge		0.07	0.093
		[2.85]	[3.57]
Other Instruments	NO	NO	YES

Table 8: 2SLS, LIML & GMM Regressions of Household Excess Portfolio Weights on Distance

This table presents the estimation results from three instrumental variables regressions of household excess portfolio weights. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $EW_{i,j} \equiv \frac{w_{i,j}-w_j^{VW}}{w_j^{VW}}$, i.e. the excess portfolio weight of household i on stock j (relative to the market value-weighted portfolio weight on stock j). The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP-Code from household i's address ZIP-Code. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. Column 1 depicts the results of the two-stage least squares regression. Column 2 depicts the results of the limited-information maximum likelihood estimation (LIML). Column 3 depicts the results of the two-step GMM estimation (GMM). The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. The (log) climate and recreation scores of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. The list of household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

Dependent Variable: Excess Portfolio Weight $EW_{i,j} \equiv \frac{w_{i,j}-w_j^{VW}}{w_j^{VW}}$			
	(1) 2SLS	(2) LIML	(3) GMM
Distance	-0.074	-0.073	-0.067
	[-1.90]	[-1.91]	[-1.95]
LogSize	-0.606	-0.606	-0.544
	[-6.23]	[-6.23]	[-6.74]
BTM	-0.368	-0.368	-0.186
	[-0.32]	[-0.32]	[-0.19]
Turnover	2.202	2.202	2.962
	[1.81]	[1.81]	[2.71]
Momentum	-1.543	-1.543	-1.407
	[-6.01]	[-6.01]	[-6.24]
Volatility	29.342	29.342	24.008
	[5.54]	[5.54]	[5.76]
Profitability	-0.58	-0.58	-0.425
	[-1.5]	[-1.5]	[-1.28]
Investment	-0.53	-0.53	-0.445
	[-3.5]	[-3.5]	[-3.59]
LogIncome	-0.042	-0.042	-0.014
	[-0.36]	[-0.36]	[-0.13]
Education	-0.629	-0.63	-0.671
	[-1.25]	[-1.25]	[-1.49]
ProfProxim	8.119	8.118	7.56
	[5.09]	[5.09]	[5.16]
Stock Industry FE	YES	YES	YES
HH Occupation-Code FE	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES
HH Location Demo's	YES	YES	YES
Stock MSA Demo's	YES	YES	YES
× HH Location Demo's	YES	YES	YES
Stock MSA Climate	YES	YES	YES
Stock MSA Recreation	YES	YES	YES

Table 9: Tobit Regressions of Household Portfolio Choice, No Correction for Location Choice

This table presents the estimation results from three Tobit regressions of household portfolio choice without correcting for location choice. The estimation is performed in a panel of households and stocks for every month separately. The depicted results are the average monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $w_{i,c,j}$, i.e. the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP-Code from household i's address ZIP-Code. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. In Column 1, the controls consist of the stock financial characteristics and the household demographics. In Column 2, demographics of the MSAs of the stocks' headquarters are added as controls. In Column 3, the (log) scores of climate and recreation of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all other demographics of the MSAs of the stocks' headquarters are also added as controls. The list of the household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes the logs of: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

	Dependent Variable: Portfolio Weight $w_{i,j}$		
	(1)	(2)	(3)
Distance	-0.014	-0.014	-0.014
	[-5.36]	[-6.06]	[-6.13]
LogSize	0.374	0.365	0.364
	[43.5]	[45.94]	[47.32]
BTM	0.052	0.049	0.05
	[6.18]	[6.06]	[6.27]
Turnover	0.687	0.63	0.541
	[12.35]	[10.5]	[11.28]
Momentum	-0.194	-0.195	-0.198
	[-8.59]	[-8.96]	[-8.99]
Volatility	3.12	2.985	2.861
	[20.81]	[19.98]	[18.43]
Profitability	-0.013	-0.026	-0.066
	[-0.21]	[-0.59]	[-1.25]
Investment	-0.079	-0.08	-0.097
	[-4.24]	[-4.3]	[-3.96]
LogIncome	0.028	0.028	0.028
	[2.76]	[2.72]	[2.63]
Education	0.112	0.109	0.106
	[2.04]	[2.05]	[2.01]
ProfProxim	0.356	0.349	0.324
	[2.5]	[2.63]	[2.27]
Stock Industry FE	YES	YES	YES
HH Occupation-Code FE	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES
HH Location Demo's	YES	YES	YES
Stock MSA Demo's	NO	YES	YES
× HH Location Demo's	NO	NO	YES
Stock MSA Climate	NO	NO	YES
Stock MSA Recreation	NO	NO	YES

Table 10: Tobit Regressions of Household Portfolio Choice With Correction for Location Choice

This table presents the estimation results from two Tobit regressions of household portfolio choice that correct for location choice using polynomial approximations of the control functions. The estimation is performed in a panel of households and stocks for every month separately. The table presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $w_{i,c,j}$, i.e. the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Distance, i.e. the distance (in degrees) of stock j's headquarters' ZIP-Code from household i's address ZIP-Code. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. The (log) climate and recreation score of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. In Column 1, the approximation of the control functions is cubic. In Column 2, the approximation of the control functions is quartic. The list of household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes the logs of: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

	Dependent Variable: Portfolio Weight $w_{i,j}$	
	(1)	(2)
Ψ's Approx.	Cubic	Quartic
Distance	-0.008	-0.008
	[-11.33]	[-14.12]
LogSize	0.361	0.36
	[53.08]	[58.04]
BTM	0.051	0.051
	[5.73]	[5.65]
Turnover	0.524	0.51
	[9.16]	[8.66]
Momentum	-0.196	-0.197
	[-8.3]	[-8.14]
Volatility	2.836	2.815
	[17.39]	[17.33]
Profitability	-0.071	-0.076
	[-1.5]	[-1.13]
Investment	-0.097	-0.099
	[-5.85]	[-5.94]
LogIncome	0.028	0.028
	[3.58]	[3.85]
Education	0.091	0.074
	[1.79]	[1.85]
ProfProxim	0.072	0.023
	[0.59]	[0.21]
Stock Industry FE	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct.	YES	YES
HH Location Demo's	YES	YES
Stock MSA Demo's	YES	YES
$\times$ HH Location Demo's	YES	YES
Stock MSA Climate	YES	YES
Stock MSA Recreation	YES	YES

Table 11: Structural Estimation of Retail Investor Portfolio Choice

This table presents the structural estimation results from three models of retail investor portfolio choice. These are heteroskedastic Tobit models in which the standard deviation of the error term equals $\gamma \exp(\delta \cdot Distance)$. The estimation is performed in a panel of households and stocks for every month separately. Panel A presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $w_{i,j}$, i.e. the portfolio weight of household i on stock j. The first key explanatory variable is Distance, i.e. the distance between household i's residential ZIP-Code and the ZIP-Code of stock j's headquarters. The second key explanatory variable is LocRes, i.e. household i's priors about the city of stock j's headquarters, measured by the Pearson residuals of the conditional logit. Stock j's financial characteristics are included as controls. The constant term and the standard deviation of the normal error term in the heteroskedastic Tobit regressions are not reported for brevity. A likelihood ratio test is performed in Columns 2 and 3 for restricted versions of the full model in Column 1. Panel B presents the average economic effect of Distance and LocRes on $w_{i,j}$ in basis points and as percentage of the mean. The sample period is from January 1991 until November 1996.

	Panel A: Coefficient Estimates		
	(1) Full Model	(2) No LocRes	(3) No Distance
Distance	-0.002	-0.004	
	[-3.04]	[-5.65]	
LocRes	0.126		0.138
	[16.75]		[15.55]
LogSize	0.372	0.373	0.373
	[48.13]	[44.65]	[47.59]
BTM	0.052	0.054	0.052
	[6.37]	[6.38]	[6.08]
Turnover	0.65	0.689	0.64
	[13.06]	[13.25]	[11.84]
Momentum	-0.191	-0.192	-0.192
	[-9.44]	[-8.7]	[-9.23]
Volatility	3.065	3.1	3.054
	[23.03]	[21.64]	[21.88]
Profitability	-0.029	-0.012	-0.033
	[-0.62]	[-0.17]	[-0.72]
Investment	-0.082	-0.079	-0.084
	[-4.68]	[-4.26]	[-4.76]
Stock Industry FE	YES	YES	YES
LR-Test vs. Full Model		2,185.5	212.48

	Panel B: Economic Effects in Basis Points and as Percentage of the Mean		
	(1) Whole Sample	(2) No Residual	(3) No Distance
Distance	-0.98	-2.25	
	-9.5%	-21.8%	
LocRes	1.5		1.68
	14.5%		16.3%

Figure 1: Geographical Distribution of Households in 57 MSAs and Russell 1000 Stocks

This figure depicts the ZIP-Code geographical coordinates of 8,688 households residing in 57 MSAs with a population of at least 750K at the end of the year 1996 and 1,193 publicly traded firms included in the Russell 1000 index during the sample period. The address ZIP-codes of households and the stocks' headquarters are converted to geographical coordinates based on the correspondence provided by the US Census Bureau. The horizontal axis is in longitude coordinates, while the vertical axis is in latitude coordinates. The blue circles indicate households, while the red squares indicate stocks. The sample period is from January 1991 to November 1996.

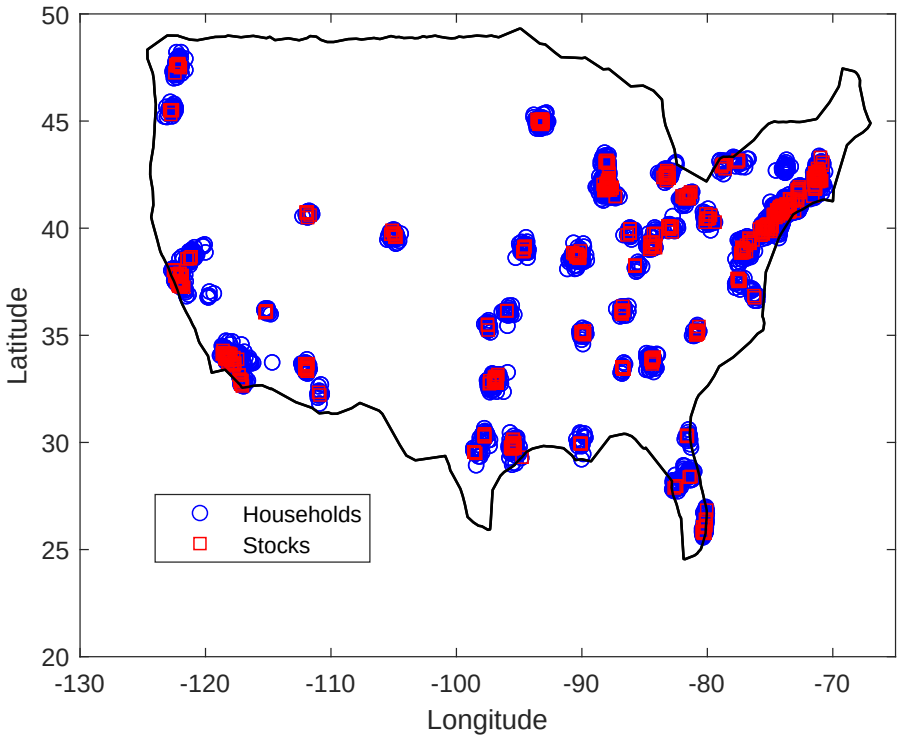


Figure 2: Coefficient Estimates of Distance Over Time in Linear Regressions of Household Excess Portfolio Weights (EW)

This figure depicts the coefficient estimates of *Distance* across the time periods from four linear excess portfolio weights regressions with stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters as controls. The vertical axis shows the values of the estimates in percentage points. The horizontal axis shows the time periods (year-month). The blue line depicts the estimates from the OLS regression which does not correct for location choice. The red dotted line depicts the estimates from an IV regression which uses as an instruments only the interaction between *LogClimate* and *LogAge*. The green dotted line depicts the estimates from an IV regression which uses as an instrument only the interaction between *LogRecreation* and *LogAge*. The yellow dotted line depicts the estimates from an GMM regression which uses as an instruments all the interactions of *LogClimate* and *LogRecreation* with the household location demographics. The sample period is from January 1991 to November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

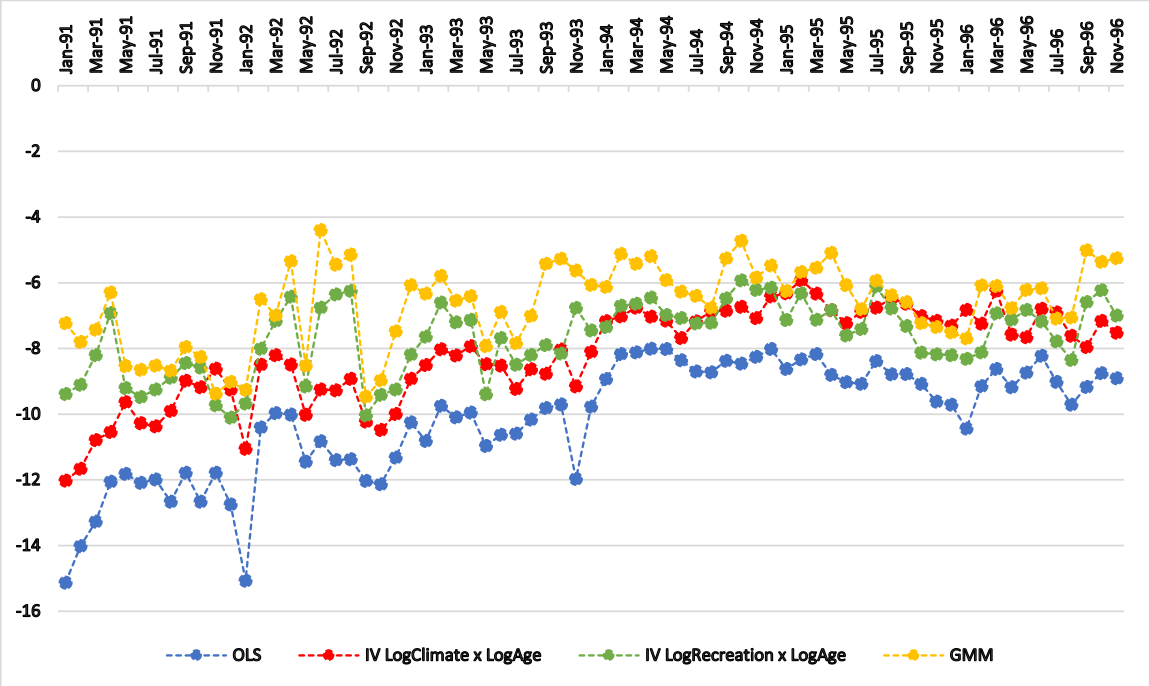
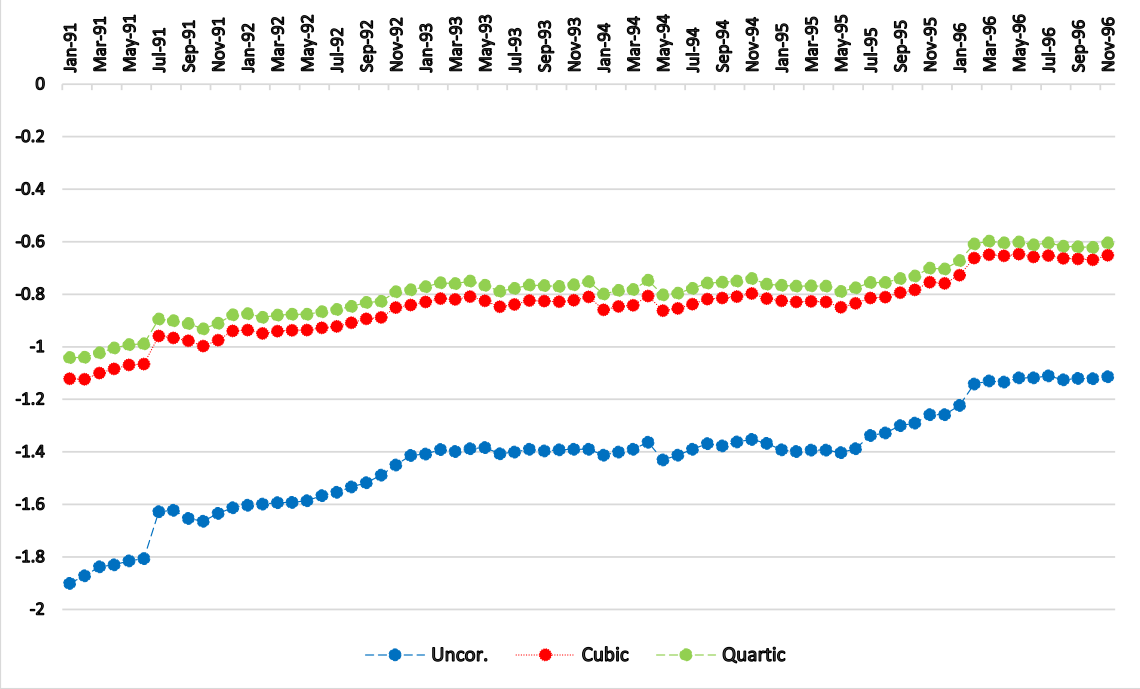


Figure 3: Coefficient Estimates of Distance Over Time in Tobit Regressions of Household Portfolio Weights

This figure depicts the coefficient estimates of *Distance* across the time periods from three Tobit portfolio choice regressions with stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters as controls. The vertical axis shows the values of the estimates in percentage points. The horizontal axis shows the time periods (year-month). The blue line depicts the estimates from the Tobit regression which does not correct for location choice. The red dotted line depicts the estimates from the Tobit regression which corrects for location choice and approximates the correction functions with third order polynomials. The green dotted line depicts the estimates from the Tobit regression which corrects for location choice and approximates the correction function with fourth order polynomials. The sample period is from January 1991 to November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.



On-line Appendix

In this On-line Appendix, we conduct a number of robustness checks and extensions. First, we consider alternative measures for the local bias, replacing the continuous distance variable measured in degrees with (i) indicators for whether the headquarters of a stock are more than a certain threshold of miles (e.g., 100 or 250 miles) away from a household's residence and (ii) the log of distance. Second, we extend our sample of stocks to the universe of Russell 3000 and our MSA sample to 80 MSAs whose population in the beginning of 1991 was at least 500,000.

Specifically, in Appendix Table 1, when we use the Tobit specification and the threshold of 100 miles, the reduction in the Away coefficient is 32% (i.e., from -0.629 in Column 1 to -0.428 in Column 2). When we use the more conservative threshold of 250 miles, the reduction in the local bias is 41% (from -0.504 in Column 3 to -0.295 in Column 4). The Tobit estimation results for the portfolio choice when we use the log of the distance are depicted in Table 2. The coefficient of $LogDist$ is decreased by 27% (from -0.164 in Column 1 to -0.119 in Column 2).

In Appendix Tables 3 and 4, we present the GMM vs. OLS estimation results for the linear under-diversification model. For the Away 100 miles dummy variable, the decrease is 28% (from -7.525 in Column 1 to -5.389 in Column 2). For the Away 250 miles dummy variable, the reductions is 32% (from -4.86 in Column 3 to -3.328 in Column 4). For $LogDist$, in Appendix Table 4, the decrease is 49.1% (from -1.671 in Column 1 to -0.85 in Column 2).

In Appendix Table 5, we depict the Tobit estimation results for households living in 80 MSAs and stocks that were members of Russell 3000 during the sample period. Without accounting for location choice, the (linear) distance coefficient is found to be -0.012. When we incorporate the control functions with the predicted location probabilities, the distance coefficient estimates decrease to -0.007. This change amounts to a 42% reduction.

In the same spirit, in Appendix Table 6, we depict the GMM estimaton results of the extended model. The average OLS distance coefficient is -0.158 (in Column 1) with an

average t-statistic of -6.2. The average GMM is is -0.083 (in Column 2) with an average t-statistic of -2.35, pointing to 47% lower local bias relative to the OLS model.

Appendix Table 1: Tobit Regressions of Household Portfolio Weights on Distance Indicator Variables

This table presents the estimation results from four Tobit regressions of household portfolio choice on distance indicator variables. The estimation is performed in a panel of households and stocks for every month separately. The table presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $w_{i,c,j}$, i.e. the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Away, i.e. an indicator variable that equals one if the distance of stock j's headquarters' ZIP Code from household i's address ZIP Code is greater than a specific threshold value. In Columns 1-2, the distance threshold is 100 miles. In Columns 3-4, the distance threshold is 250 miles. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. Columns 1 and 3 depict the results without correcting for location choice. Columns 2 and 4 depict the results correcting for location choice. The control functions are approximated with fourth order polynomials. The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. The (log) climate and recreation score of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. The list of the household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

Dependent Variable: Portfolio Weight $w_{i,j}$				
	100 Miles Away		250 Miles Away	
	Uncor. (1)	Corrected (2)	Uncor. (3)	Corrected (4)
Away	-0.629	-0.428	-0.504	-0.295
	[-5.49]	[-6.73]	[-5.04]	[-7.35]
LogSize	0.362	0.36	0.363	0.36
	[46]	[47.07]	[49.16]	[50.76]
BTM	0.047	0.05	0.051	0.052
	[3.34]	[2.96]	[3.65]	[4.08]
Turnover	0.5	0.494	0.525	0.506
	[11.12]	[8.26]	[10.24]	[8.64]
Momentum	-0.197	-0.196	-0.196	-0.196
	[-5.62]	[-5.92]	[-5.68]	[-5.92]
Volatility	2.847	2.807	2.821	2.795
	[4.29]	[5.22]	[3.91]	[5.79]
Profitability	-0.079	-0.081	-0.07	-0.076
	[-7.45]	[-8.4]	[-4.14]	[-5.54]
Investment	-0.099	-0.1	-0.099	-0.1
	[-4.12]	[-4.3]	[-3.73]	[-4.85]
LogIncome	0.023	0.026	0.023	0.026
	[2.7]	[3.42]	[2.82]	[3.27]
Education	0.056	0.074	0.057	0.065
	[1.91]	[2.09]	[1.93]	[2.04]
ProfProxim	0.196	0.03	0.283	0.023
	[2.11]	[0.28]	[2.64]	[0.37]
Stock Industry FE	YES	YES	YES	YES
HH Occupation-Code FE	YES	YES	YES	YES
HH ZIP-Code Race Pct	YES	YES	YES	YES
HH Location Demo's	YES	YES	YES	YES
Stock MSA Demo's	YES	YES	YES	YES
× HH Location Demo's	YES	YES	YES	YES
Stock MSA Climate	YES	YES	YES	YES
Stock MSA Recreation	YES	YES	YES	YES

Appendix Table 2: Tobit Regressions of Household Portfolio Weights on Log Distance

This table presents the estimation results from two Tobit regressions of household portfolio choice on distance indicator variables. The estimation is performed in a panel of households and stocks for every month separately. The table presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $w_{i,c,j}$, i.e. the portfolio weight of household i residing in city c on stock j. The key explanatory variable is LogDist, i.e. the log of the distance of stock j's headquarters' ZIP Code from household i's address ZIP Code. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. Column 1 depicts the results without correcting for location choice. Column 2 depicts the results correcting for location choice. The control functions are approximated with fourth order polynomials. The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. The (log) climate and recreation score of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. The list of the household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

Dependent Variable: Portfolio Weight $w_{i,j}$		
	(1) Uncor.	(2) Corrected
LogDist	-0.164 [-6.34]	-0.119 [-9.49]
LogSize	0.363 [48.46]	0.36 [53.64]
BTM	0.049 [4.37]	0.051 [4.96]
Turnover	0.522 [11.01]	0.508 [9.94]
Momentum	-0.196 [-8.59]	-0.196 [-8.22]
Volatility	2.851 [10.8]	2.805 [7.71]
Profitability	-0.074 [-1.4]	-0.077 [-1.52]
Investment	-0.098 [-3.19]	-0.099 [-5]
LogIncome	0.025 [2.81]	0.027 [3.22]
Education	0.027 [0.5]	0.044 [6.12]
ProfProxim	0.096 [2.01]	0.008 [0.06]
Stock Industry FE	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct	YES	YES
HH Location Demo's	YES	YES
Stock MSA Demo's	YES	YES
× HH Location Demo's	YES	YES
Stock MSA Climate	YES	YES
Stock MSA Recreation	YES	YES

Appendix Table 3: OLS and GMM Regressions of Household Excess Portfolio Weights on Distance Indicator Variables

This table presents the estimation results from four linear regressions of household excess portfolio weights on distance indicator variables. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $EW_{i,j} \equiv (w_{i,j} - w_j^{VW})/w_j^{VW}$, i.e. the excess portfolio weight of household i on stock j (w.r.t. to the market value-weighted portfolio weight on stock j). The key explanatory variable is Away, i.e. an indicator variable that equals one if the distance of stock j's headquarters' ZIP Code from household i's address ZIP Code is greater than a specific threshold value. In Columns 1-2, the distance threshold is 100 miles. In Columns 3-4, the distance threshold is 250 miles. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. Columns 1 and 3 depict the ordinary least squares regression results. Columns 2 and 4 depict the two-step GMM estimation results. The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. The F-statistic tests the hypothesis that the coefficients of the instruments are jointly zero - taking into account the two-level clustering. The Hansen J-test is the test of overidentifying restrictions. The (log) climate and recreation score of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. The list of household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

Dependent Variable: Excess Portfolio Weight $EW_{i,j} \equiv (w_{i,j} - w_j^{VW})/w_j^{VW}$				
	100 Miles Away		250 Miles Away	
	(1) OLS	(2) GMM	(3) OLS	(4) GMM
Away	-7.525	-5.389	-4.86	-3.328
	[-3.98]	[-1.7]	[-4.84]	[-1.82]
LogSize	-0.613	-0.545	-0.611	-0.548
	[-6.08]	[-6.71]	[-6.02]	[-6.81]
BTM	-0.396	-0.109	-0.377	-0.12
	[-0.38]	[-0.18]	[-0.34]	[-0.14]
Turnover	1.857	2.895	2.116	2.861
	[1.56]	[2.46]	[1.78]	[2.65]
Momentum	-1.536	-1.403	-1.541	-1.398
	[-5.64]	[-6.19]	[-5.65]	[-6.21]
Volatility	29.333	24.301	29.257	24.15
	[5.24]	[5.8]	[5.24]	[5.79]
Profitability	-0.621	-0.427	-0.549	-0.407
	[-1.24]	[-1.26]	[-1.06]	[-1.2]
Investment	-0.539	-0.461	-0.541	-0.459
	[-3.51]	[-3.7]	[-3.51]	[-3.74]
LogIncome	-0.092	-0.034	-0.09	-0.061
	[-0.84]	[-0.27]	[-0.81]	[-0.5]
Education	-1.029	-0.781	-1.018	-1.076
	[-2.35]	[-1.12]	[-2.24]	[-1.44]
ProfProxim	7.358	6.714	7.918	7.438
	[4.11]	[4.74]	[4.52]	[5.37]
Stock Industry FE	YES	YES	YES	YES
HH Occupation-Code FE	YES	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES	YES
HH Location Demo's	YES	YES	YES	YES
Stock MSA Demo's	YES	YES	YES	YES
× HH Location Demo's	YES	YES	YES	YES
Stock MSA Climate	YES	YES	YES	YES
Stock MSA Recreation	YES	YES	YES	YES
First-stage F-statistic		25.42		23.01
Hansen J-test (p-value)		0.69		0.73

Appendix Table 4: OLS and GMM Regressions of Household Excess Portfolio Weights on Log Distance

This table presents the estimation results from two linear regressions of household excess portfolio weights on distance indicator variables. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $EW_{i,j} \equiv (w_{i,j} - w_j^{VW})/w_j^{VW}$, i.e. the excess portfolio weight of household i on stock j (w.r.t. to the market value-weighted portfolio weight on stock j). The key explanatory variable is LogDist, i.e. the log of the distance of stock j's headquarters' ZIP Code from household i's address ZIP Code. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. Columns 1 depicts the ordinary least squares regression results. Column 2 depicts the two-step GMM estimation results. The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. The F-statistic tests the hypothesis that the coefficients of the instruments are jointly zero - taking into account the two-level clustering. The Hansen J-test is the test of overidentifying restrictions. The (log) climate and recreation score of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. The list of household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 57 MSAs with a population of at least 750K at the end of 1996. The investment universe consists of Russell 1000 stocks headquartered in these MSAs.

Dependent Variable: Excess Portfolio Weight $EW_{i,j} \equiv (w_{i,j} - w_j^{VW})/w_j^{VW}$		
	(1) OLS	(2) GMM
LogDist	-1.671 [-5.21]	-0.85 [-1.89]
LogSize	-0.609 [-6.03]	-0.545 [-6.81]
BTM	-0.383 [-0.36]	-0.115 [0.15]
Turnover	2.147 [1.82]	2.864 [2.65]
Momentum	-1.536 [-5.63]	-1.4 [-6.27]
Volatility	29.419 [5.28]	24.243 [5.83]
Profitability	-0.597 [-1.18]	-0.448 [-1.33]
Investment	-0.531 [-3.42]	-0.451 [-3.68]
LogIncome	-0.062 [-0.55]	-0.033 [-0.29]
Education	-0.967 [-1.83]	-0.826 [-1.71]
ProfProxim	7.47 [4.13]	7.475 [5.2]
Stock Industry FE	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct.	YES	YES
HH Location Demo's	YES	YES
Stock MSA Demo's	YES	YES
× HH Location Demo's	YES	YES
Stock MSA Climate	YES	YES
Stock MSA Recreation	YES	YES
First-stage F-statistic		26.41
Hansen J-test (p-value)		0.58

Appendix Table 5: Tobit Regressions of Household Portfolio Weights on Distance for Russell 3000 and 80 MSAs

This table presents the estimation results from two Tobit regressions of household portfolio choice. The estimation is performed in a panel of households and stocks for every month separately. The table presents the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $w_{i,c,j}$, i.e. the portfolio weight of household i residing in city c on stock j. The key explanatory variable is Distance, i.e. the distance of stock j's headquarters' ZIP Code from household i's address ZIP Code. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. The (log) climate and recreation score of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. The list of household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 80 MSAs with a population of at least 500K in the beginning of 1991. The investment universe consists of Russell 3000 stocks headquartered in these MSAs.

Dependent Variable: Portfolio Weight $w_{i,j}$		
	(1) Uncor.	(2) Corrected
Distance	-0.012 [-8.28]	-0.007 [-15.28]
LogSize	0.262 [56.72]	0.26 [52.02]
BTM	0.028 [1.47]	0.028 [1.24]
Turnover	0.354 [11.06]	0.331 [10.72]
Momentum	-0.071 [-4.42]	-0.07 [-3.99]
Volatility	1.375 [18.63]	1.358 [19.86]
Profitability	-0.025 [-1.06]	-0.031 [-1.47]
Investment	-0.047 [-4.19]	-0.048 [-3.42]
LogIncome	0.029 [5.29]	0.028 [3.23]
Education	0.108 [2.57]	0.074 [1.93]
ProfProxim	0.333 [4.3]	0.01 [0.6]
Stock Industry FE	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct.	YES	YES
HH Location Demo's	YES	YES
Stock MSA Demo's	YES	YES
× HH Location Demo's	YES	YES
Stock MSA Climate	YES	YES
Stock MSA Recreation	YES	YES

Appendix Table 6: OLS and GMM Regressions of Household Excess Portfolio Weights on Distance for Russell 3000 and 80 MSAs

This table presents the estimation results from two linear regressions of household excess portfolio weights. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the averages of the monthly coefficient estimates and the average t-statistics [shown in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. The dependent variable is $EW_{i,j} \equiv (w_{i,j} - w_j^{VW})/w_j^{VW}$, i.e. the excess portfolio weight of household i on stock j (w.r.t. to the market value-weighted portfolio weight on stock j). The key explanatory variable is Distance, i.e. the distance of stock j's headquarters' ZIP Code from household i's address ZIP Code. Stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters are included as controls. See Table 1 for a detailed description. Column 1 depicts the ordinary least squares regression results. Column 2 depicts the two-step GMM estimation results. The instruments are pair-wise interaction terms based on the (log) score of climate and recreation of the MSAs of the stocks' headquarters and the household location demographics. The F-statistic tests the hypothesis that the coefficients of the instruments are jointly zero - taking into account the two-level clustering. The Hansen J-test is the test of overidentifying restrictions. The (log) climate and recreation score of the MSAs of the stocks' headquarters and the interactions between the household location demographics and all *other* demographics of the MSAs of the stocks' headquarters are added as controls. The list of household location demographics consists of: the log of the age of the head, the log of the family size, an indicator variable equal to one if the head is married and an indicator variable equal to one if the head is male. The list of the demographics of the MSAs of the stocks' headquarters includes in logs: the income per capita, the unemployment rate, the house price index, the population, the score of transportation, the score of colleges, the score of crime and the score of health care. The sample period is from January 1991 until November 1996. Households reside in 80 MSAs with a population of at least 500K in the beginning of 1991. The investment universe consists of Russell 3000 stocks headquartered in these MSAs.

Dependent Variable: Excess Portfolio Weight $EW_{i,j} \equiv (w_{i,j} - w_j^{VW})/w_j^{VW}$		
	(1) OLS	(2) GMM
Distance	-0.158 [-6.2]	-0.083 [-2.35]
LogSize	-1.064 [-7.69]	-0.957 [-9.63]
BTM	0.191 [0.63]	0.115 [0.58]
Turnover	2.97 [2.05]	2.674 [2.4]
Momentum	-0.558 [-3]	-0.484 [-3.35]
Volatility	16.005 [3.13]	15.732 [3.84]
Profitability	0.003 [0.29]	0.005 [0.26]
Investment	-0.26 [-1.97]	-0.178 [-1.95]
LogIncome	-0.08 [-0.24]	-0.061 [-0.35]
Education	-0.475 [-0.38]	-0.631 [-0.73]
ProfProxim	6.614 [2.71]	5.223 [2.99]
Stock Industry FE	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct.	YES	YES
HH Location Demo's	YES	YES
Stock MSA Demo's	YES	YES
× HH Location Demo's	YES	YES
Stock MSA Climate	YES	YES
Stock MSA Recreation	YES	YES
First-stage F-statistic		24.39
Hansen J-test (p-value)		0.74