

Location Choice, Portfolio Choice*

Ioannis Branikas[†] Harrison Hong[‡] Jiangmin Xu[§]

First Draft: September 2016

This Draft: November 2018

Abstract

Households hold undiversified stock portfolios of firms headquartered near their city of residence. Leading explanations 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 use location-choice models to account for this selection. We propose as instruments that older households prefer to locate in areas with recreation and mild climate for non-pecuniary reasons. Our analysis yields causal estimates for proximity and points to Pearson residuals from location-choice models as measures of latent household expectations.

*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, Kai Li, Philip Bond, Will Gornall 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, the 2017 Helsinki Finance Summit and the 2018 Pacific Northwest Finance Conference 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 stock-portfolio weights on demographic and stock characteristics, distance from household residence to firm headquarters emerges as a key explanatory variable. Yet, investors do not get rewarded for holding such concentrated local stock portfolios. 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, a number of theories have been given for this local bias in the literature.² Leading explanations assign a causal role for proximity. One 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 — be it the company they work for, companies near them that they know friends at, or even the telephone company that services their homes — because they feel competent in evaluating them. Another causal explanation for local bias is Keeping up with the Joneses’ preferences ([Luttmer \(2005\)](#), [Charles, Hurst, and Roussanov \(2009\)](#)), which leads to a demand for local

¹Prominent studies include Finland ([Grinblatt and Keloharju \(2001\)](#)), US ([Zhu \(2002\)](#)), and China ([Feng and Seasholes \(2008\)](#)) to name a few. 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, as would be needed in a rational information acquisition story (see, e.g., [Seasholes and Zhu \(2010\)](#)). 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.

²There is a sizable 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 since local stock picks do not significantly outperform the market.

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

Regardless of the particular explanation, 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 (McFadden (1978)), agents optimally locate in cities that provide them with the highest utility. 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 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 heterogeneous enough subjective expectations. Such optimal spatial-sorting models are consistent with migration patterns that one sees in the US (e.g., Diamond (2016), Kaplan and Schulhofer-Wohl (2017)).

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 locate to areas which they view as having a bright economic future, not only for themselves but for their family and next 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. This optimism can generate speculative purchases of local stocks or hedging demand, since optimistic households with Keeping-up-with-the-Jones preferences might naturally also expect greater run-ups in scarce local resources.

More importantly, their latent expectations about cities that they did not locate 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 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 by utilizing location-choice models from urban economics. We apply this methodology to a sample of household portfolios 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.

We consider two widely used reduced-form portfolio specifications: a non-linear Tobit model where the dependent variable is portfolio weights and a linear model where the dependent variable is household portfolio deviations from a market benchmark. 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 endogenous location decisions or selection on the local bias of portfolio choices, we estimate a location choice model for our sample. 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.

We propose as instruments that older households 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). This is a robust empirical finding in the urban economics (see, e.g., [Sinha, Caulkins, and Cropper \(2017\)](#)), reflecting a non-pecuniary motive for location due to either biological reasons or that the head of the household might be close to retirement. We find that the interactions of the age of the household with the recreation and climate scores of a MSA are among the strongest predictors of household location choice in our sample.

We discuss in detail the strengths and weaknesses of our identification strategy in Section 3.4. First, balance tests ([Roberts and Whited \(2013\)](#)) 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 predicted by the climate or recreation score in a statistically significant way. Second, our instruments have to be uncorrelated with unobservables in the second-stage portfolio weights regression. That is, 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 Dallas - Fort Worth 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.

Interactions of age and MSA recreation and climate ought to be excluded in many tra-

ditional models of household portfolios. However, a major concern for us is a non-standard model where older households read fishing magazines geared toward vacation, while younger households read magazines about music festivals. If households bought stocks based on these magazines, and vacation services companies are more likely to be located in Miami while music festival companies are headquartered in Dallas, we would violate the exclusion restriction.

It is not obvious that this mechanism is economically large, because if it were, we might not see much local bias to begin with. Nonetheless, we can be conservative and drop consumer-oriented companies or companies with a high advertising expenditure. In this subsample of stocks, our exclusion restriction ought to be even more sound. We also propose several other instruments using other household demographics interacted with climate and recreation. We find our results to be robust.

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. 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.³ We get a large correction of 43% in the distance coefficient.

To correct the endogeneity of distance in the linear deviations model, we can use standard instrumental-variables regression approaches. 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. While the non-linear model is a better model of household portfolios,

³In the Online Appendix, we show that our results apply also for the extended investment universe of Russell 3000.

this linear analysis is also useful since it shows that the corrections in the non-linear model is not an artifact of non-linearities in the correction function.

The main contribution of our paper is to introduce location choice models into the analysis of household portfolio choice. Beyond providing a better regression framework for modeling the causal role of proximity for portfolios, our analysis also points to using location-choice model residuals as noisy measures of latent household expectations regarding economic vibrancy of MSAs. In the last part of our paper, we develop these noisy measures and show how they might be integrated into and expand traditional tools in household portfolio analysis. We also show that they have surprisingly large explanatory power for these portfolios.

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.⁴

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 variables are observed at an annual frequency.

In addition, we contribute to the literature of location choice by collecting additional variables which 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

⁴In 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.

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, (v) the supply of recreational assets, and (vi) climate mildness. The analytical definition of these variables is given in the Online 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 Panel A of Table 1.⁵ 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

⁵The mean income per capita is 21.6 thousand dollars. 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.

refer to 401(k) plans. Therefore, mechanical effects on the stock choice from the shares of an employer are limited.

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 \(2006\)](#).⁶ Requiring the complete observability of these additional household demographics leaves us 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.⁷ 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.⁸ Their first time-series observations comprise the sample of our location choice model. The summary statistics of the household

⁶For 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.

⁷In the Online Appendix, when we focus on 80 MSAs, we have 10,261 households.

⁸According to the US Census Bureau, the average percentage of movers during our sample period (1991-1996) was 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.

demographics are presented in Panel B of Table 1.⁹

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.¹⁰ 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. 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 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 $t - 1$.¹¹ We also use the Fama-French industry classification of

⁹The 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 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%.

¹⁰Since 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.

¹¹As 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 July of year t until

stocks into 17 categories based on the four-digit SIC code, which is available from Kenneth R. French's website.¹² We depict the summary statistics of the stock financial characteristics in Panel C of Table 1.¹³

2.4. Household Stock Holdings

The portfolio positions of households are summarized in Panel D 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.¹⁴ 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.¹⁵

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 0.03.¹⁶

June of year $t + 1$.

¹²<http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html>

¹³ 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 Russell 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).

¹⁴According to the 2003 US Census Bureau report on net worth and asset ownership of households, the median value of stockholdings for a typical US household in 1998 was \$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.

¹⁵For all households and all stocks in the database, Barber and Odean (2000) document an average annual portfolio turnover of 75%.

¹⁶These figures are very similar to the portfolio summary statistics reported by Gargano and Rossi (2018)

2.5. Local Bias of Stock Portfolios

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.¹⁷ The geographical distribution of the 8,688 households in our sample and the 1,193 stocks in Russell 1000 is presented in Appendix 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 summary statistics of the local bias (LB) in our household stock holdings data are given in Panel E of Table 1 and are constructed as in Coval and Moskowitz (1999). Column 1 (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 household i 's residential area and the headquarters' area of 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 2 (labeled "Avg. Distance from Benchmark") reports the average portfolio weighted distance of households from a 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, while Row 2 refers to the value-weighted portfolio. Column 3 (labeled "Difference") reports the average difference between Column 2 and Column 1, which is essentially the average local bias of households in distance units. Column 4 (labeled "% Bias (LB)") reports the local bias (LB) measure as a percentage of the benchmark distance.¹⁸ Column 5 reports the t -statistics for the LB measure.

for a recent broker during the period January 2013 to June 2014.

¹⁷We measure distance in degrees. Multiplying by $2\pi R/360$ converts it to miles (kilometers), where $R \approx 3,963$ miles (6,378 kilometers).

¹⁸Column 3 shows $\frac{1}{I} \sum_{i=1}^I \sum_{j=1}^J (w_{i,j} - \bar{w}_j) dist_{i,j}$ and Column 4 shows $\frac{1}{I} \sum_{i=1}^I \left(\sum_{j=1}^J (w_{i,j} - \bar{w}_j) dist_{i,j} / \sum_{j=1}^J \bar{w}_j dist_{i,j} \right)$.

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% of the benchmark, while, using the value-weighted portfolio, it is slightly decreased to 8.26 or 43.72% of the benchmark.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 $\ell = 1, \dots, C$ 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:

$$c = \arg \max_{\ell \in \{1, \dots, C\}} \{V_{i,\ell} + e_{i,\ell}\} \quad (1)$$

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 $K \times 1$

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

vector \mathbf{z}_ℓ . 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.:

$$V_{i,\ell} = \boldsymbol{\rho}_i \mathbf{z}_\ell \quad (2)$$

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, \mathbf{D}_i , (through a $K \times M$ matrix of parameters $\boldsymbol{\Pi}$), i.e.:

$$\boldsymbol{\rho}_i = \boldsymbol{\rho} + \boldsymbol{\Pi} \mathbf{D}_i \quad (3)$$

The vector of household i 's demographics, \mathbf{D}_i , that we use in our empirical analysis has as elements its family size and the age, gender and marital status of its head. We refer to these variables as "Household Basic Demographics".²¹ By combining Equations (2) and (3), we eventually represent household i 's observed utility from locating in city ℓ as:

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

where δ_ℓ is the observed utility from the characteristics of city ℓ that is common for all

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

²¹We 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.

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 $\boldsymbol{\theta}^{loc} \equiv (\boldsymbol{\rho}, \mathbf{\Pi})$ from the data, we will have also estimated the observed utilities of household i from all the available locations, $\{V_{i,\ell}\}_{\ell=1,\dots,C}$.

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

$$v_{i,c} = \max_{\ell \in \{1,\dots,C\}/c} \{V_{i,\ell} - V_{i,c} + e_{i,\ell} - e_{i,c}\} \quad (5)$$

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

$$r_{i,c} = \mathbf{1} [v_{i,c} < 0] \quad (6)$$

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:

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

3.2. Reduced-Form Portfolio Choice Regressions

We consider two widely-used regression specifications in the literature. 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. Specifically, consistent with the fact that households do not short, we first assume that household i , residing in city c , decides how much to invest in stock j , headquartered in city h , according to a linear factor rule censored at zero:

$$w_{i,c,h,j} = (\alpha + \boldsymbol{\beta}\mathbf{x}_j + \boldsymbol{\gamma}\mathbf{D}_i + \delta dist_{i,c,h,j} + \epsilon_{i,c,h,j})^+ \quad (8)$$

where $(\cdot)^+ \equiv \max\{\cdot, 0\}$ captures both household i 's extensive and intensive margin. \mathbf{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. \mathbf{D}_i is the vector of household i 's demographics, as in its location choice problem. 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, $\epsilon_{i,c,h,j}$ 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 above empirical censored specification is appropriate for capturing the high degree of sparsity that household portfolios exhibit (e.g., including on average only two stocks in Russell 1000). 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

For tractability reasons, the literature (e.g., [Goetzmann and Kumar \(2008\)](#), [Brandt, Santa-Clara, and Valkanov \(2009\)](#)) has also 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 second model $\boldsymbol{\theta}^{dev} \equiv (\alpha^{dev}, \boldsymbol{\beta}^{dev}, \boldsymbol{\gamma}^{dev}, \delta^{dev})$ and estimate them by running the following linear regression for every month in the sample period:

$$\frac{w_{i,c,h,j} - w_j^{VW}}{w_j^{VW}} = \alpha^{dev} + \boldsymbol{\beta}^{dev} \mathbf{x}_j + \boldsymbol{\gamma}^{dev} \mathbf{D}_i + \delta^{dev} dist_{i,c,h,j} + \epsilon_{i,c,h,j}^{dev} \quad (9)$$

²²The 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 \(2006\)](#).

²³In 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).

where 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. The caveat of this specification is that the many zero portfolio weights on stocks, that motivated the use of the Tobit model in the first place, are translated to many 100% negative deviations from the market.

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:

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

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 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:

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

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 (e.g. latent demand for stock) when it lives in city c and considers investing in stock j headquartered in city h , is correlated with the idiosyncratic location choice (e.g. latent expectations about prospects of location), $\{e_{i,\ell}\}_{\ell=1}^C$ - 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, $v_{i,c}$. That is, latent optimism about a location's economic prospects is going to lead to latent speculative or hedging demand for local stocks in that city, since they should load on the local economy to some degree. To show such a potential correlation more clearly, we decompose $\epsilon_{i,c,h,j}$ as follows:

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

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:

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

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 prin-

ciple 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 (8) and (9).²⁵

3.4. Identification Strategies

3.4.1. Instruments

We propose as instruments that older households, which for biological reasons or reasons having to do with 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 that our climate and recreation scores have low correlations with other MSA observables that could be relevant for pecuniary motives. In Panels A and B of Table 2, we split

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

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

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

²⁵Since 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 expectations for an area (here, captured by the control function) are positively correlated with unobservable latent investment demand for stocks headquartered there, we anticipate a *local bias over-estimation* when location choice is ignored.

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 Panel C of Table 2, 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 between the age of the household and the recreation or climate score in the MSA where the stock is headquartered, while controlling for household and MSA demographics.

This exclusion restriction might be violated under certain conditions: (1) older households

²⁶The 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.

read fishing magazines (e.g. American Angler) geared toward vacation while younger households read magazines about music festivals (e.g. Billboard), (2) vacation services companies are more likely to be located in Florida while music festival companies are headquartered in Dallas, and (3) households buy based on advertisements in these magazines. It is not clear ex-ante how plausible this mechanism is, since it would have households potentially buying stocks far away from them, in contradiction of the observed local bias in the data. Nonetheless, we can address this by conducting a robustness analysis in which we drop stocks of consumer-oriented companies or companies with high advertising expenditure. In this subsample of stocks, this alternative mechanism ought not to be relevant.

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. It does not matter if locational decisions can 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.

3.4.2. Control Function Approach in the Non-linear Model

In the non-linear model of Equation (8), 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:

$$\psi_{c,h}(\{V_{i,\ell}\}_{\ell=1}^C) = \Psi_{c,h}(\{p_{i,\ell}\}_{\ell=1}^C) \quad (14)$$

Based on Equation (14), 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. Equations (8), (12), (13) and (14) yield that the portfolio weights regression that corrects for location choice is:

$$w_{i,c,h,j} = \left(\alpha + \beta \mathbf{x}_j + \gamma \mathbf{D}_i + \delta \text{dist}_{i,c,h,j} + \Psi_{c,h} \left(\{p_{i,\ell}\}_{\ell=1}^C \right) + \eta_{i,c,h,j} \right)^+ \quad (15)$$

3.4.3. GMM Approach in the Linear Model

In the linear model of Equation (9), we directly examine the effect of the instruments on distance using the following first-stage equation:

$$d_{i,c,h,j} = \iota + \kappa \mathbf{x}_j + \lambda \mathbf{D}_i + \phi \mathbf{D}_i \times \mathbf{z}_h + \omega_{i,c,h,j} \quad (16)$$

where the interactions of household i ' demographics with the climate and recreation scores 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.

3.5. What about Movers?

As we have noted, there are no movers in our short sample period. But having movers would not solve any endogeneity issues. If we saw that households changed their local portfolios

when they moved, we would still have to address the same endogeneity issues. Why are they moving and what were their latent expectations? That is, rather than estimating a static location choice model, one would have to estimate a dynamic location choice model. But we would end up having to use the same instruments.

Moreover, the fact that there might not be much moving per se (since many people are likely to live near their place of birth) is also not an issue for our sample to the extent households had a choice to move in the first place. As we show below and is known in the literature, there are large residuals in our location-choice models which reflect sorting due to unobserved heterogeneities.

4. Estimation

4.1. Location Choice

In Table 3, we present the conditional logit estimation of three location choice models. Across all three columns, the dependent variable is the indicator variable $r_{i,c}$ from Equation (6), which equals one if household i resides in MSA c . Since we have 8,688 unique households living in 57 different MSAs, the sample consists of 495,216 observations. As the results of these regressions are well known, we focus on our independent variables of interest $\text{LogClimate} \times \text{LogAge}$ and $\text{LogRecreation} \times \text{LogAge}$ in Columns 1 and 2. In Column 3, we also include in the specification LogClimate and LogRecreation interacted with other Household Basic Demographics, i.e. LogFamSize , Male and Married . Across all three columns, we control for other MSA demographics besides climate and recreation (listed in Panel A of Table 1) and interactions of these with Household Basic Demographics.

In Column 1, the coefficient estimate on $\text{LogClimate} \times \text{LogAge}$ is 0.459 and has a t -statistic of 4.02 — which is among the highest in the logit model. That is, older households are more likely to live in mild climate MSAs. There are some other significant interactions for location choice which are not shown for brevity. For example, one of them is $\text{LogHPI} \times$

LogFamSize, since large families are more likely move to cities with lower house prices. In Column 2, we turn our attention to the instrument $LogRecreation \times LogAge$. In the same spirit, older households prefer to locate closer to MSAs with recreation. The estimated coefficient of the instrument is 0.633 and has a t -statistic of 4.67.

In terms of economic effects, 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 among the strongest predictors of household location choice in our sample.

Recall that other household demographics, besides age, can also affect the propensity of a household to have a non-pecuniary motive for location — as we discuss above — and hence might be useful as additional instruments. To this end, in Column 3, In Column 3, we include both $LogClimate \times LogAge$ and $LogRecreation \times LogAge$ as well as the interactions of the MSA climate and recreation scores with the other Household Basic Demographics. The coefficients on our two primary variables of interest remain similar. Moreover, the other interactions, particular the ones referring to family size, also do explain location choice. Although not shown for brevity since there are many interactions, the t -statistics of $LogClimate \times LogFamSize$ and $LogRecreation \times LogFamSize$ are respectively -2.3 and -2.43 .

We also conduct likelihood ratio tests to compare the fit of the full model in Column 3 against Columns 1 and 2. We reject the restricted versions at any reasonable level of statistical significance. The LR statistic against the model in Column 1 (with only $LogClimate \times LogAge$ as instrument) is 126.97, while against the model in Column 2 (with only $LogRecreation \times LogAge$ as instrument) is 75.15. Toward a similar end, the Akaike information criterion in Column 3 is also the lowest one, pointing to it being the better model ([Akaike \(1974\)](#)). As such, we will use Column 3 as our preferred first-stage regression for location choice.

4.2. Selection Model for Portfolio Choice and Distance

We now estimate the nonlinear Tobit model for portfolio weights as defined in Equation (15), where we model the selection of location choice using the location choice models in Table 3. As it 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 :

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

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

$$\Psi_{c,h}(p_{i,c}, p_{i,h}) = \begin{cases} \Psi_c(p_{i,c}) & \text{if } h = c \\ \Psi_h(p_{i,c}, p_{i,h}) & \text{if } h \neq c \end{cases} \quad (18)$$

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

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

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. $\Psi^s(\cdot)$ for the case in which household i considers investing in a stock that is headquartered in the same city in which it resides and $\Psi^d(\cdot)$ for the case in which the stock's headquarters are located in a different city.²⁷ Both $\Psi^s(\cdot)$ and $\Psi^d(\cdot)$ can be flexibly estimated through a polynomial series expansion.

In Table 4, we present the Tobit estimation results. We estimate the models separately for every month in our sample 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. Our sample is an unbalanced panel of 8,688 households

²⁷In 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:

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

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

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

and 1,193 Russell 1000 stocks for the whole period. On average, in every month, we have 4,339 households choosing from 988 stocks.

In Column 1 of Table 4, which shows the Tobit model without controlling for location choice (as it is widely used in the household finance literature), it is easy to see that distance is among the most powerful predictors of portfolio weights.

The uncorrected distance coefficient estimate is on average -0.014 , with a corresponding average t -statistic of -6 . The sign and magnitude of the estimated coefficients of the other controls are also as anticipated. For instance, size, book-to-market ratio, turnover and volatility all have positive and statistically significant coefficients.

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*. 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 Panel E of Table 1, *Distance* is estimated to have a very sizable economic effect on the household portfolio choice.

In Columns 2 and 3, we use the specification from Column 3 of Table 3 as our model of location choice to address the issue of selection bias in the second-stage portfolio weight regression. Recall that this model is estimated using all 8,688 households in our sample. We use this model to then generate the location choice probabilities for the cross-section of households (on average 4,339 of them) in any given month. These location probabilities are then introduced as arguments of the control functions to adjust for selection in that month. We exclude the interactions between the climate and recreation scores of the MSAs of the stocks' headquarters and the Basic Household Demographics. These are our instruments and affect household stock-investment only through the location probabilities of the control

functions.

Best practice in the literature is that a linear or quadratic approximation of the control functions is not that flexible (e.g., Bourguignon, Fournier, and Gurgand (2007)). As such, in Column 2, we use a cubic approximation for the control functions, while, in Column 3, the approximation is quartic. Regardless of the polynomial approximation used, 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. Interestingly, the coefficients have higher t -statistics than the uncorrected results in Column 1. This is possible to the extent the polynomial correction functions soak up variation and improve the fit of the overall model. We also try a fifth and sixth order polynomial approximation, without finding any further significant decrease in the distance coefficient.

Subfigure 1a depicts the *Distance* coefficient estimates (multiplied by 100) in the Tobit 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. There is a dramatic difference between the uncorrected estimates and the causal estimates across all the months.

4.3. Linear IV for Portfolio Under-Diversification and Distance

One worry always with non-linear models and the control function approach is that non-linearities might be driving the results. In other words, adding non-linear functions as control terms could inadvertently influence the coefficient on distance. Even though this non-linear model better captures household behavior than the linear deviations in weights specification, it is nonetheless valuable to see how the results look in this linear set-up. This set-up also allows us to apply well-known instrumental variables techniques not available for non-linear models such as tests for weak instruments. Our instrumental variables analysis follows closely the guidelines of Angrist and Pischke (2008), where the idea is that if one has multiple

instruments, the key is to show robustness across a variety of regression specifications.

To this end, in Table 5, we estimate in Panel A the first-stage regression of the distance of a household from the headquarters of a given stock in Equation (16) on the instruments of interest, and in Panel B we show how the corresponding second-stage regressions involving portfolio deviations ((9)) differ from OLS as we use different instruments.

In Panel A, the interactions of *LogAge* with *LogClimate* and *LogRecreation*, which said that older households prefer to live in mild climate and high recreation MSAs, will now equivalently predict a shorter distance between an older household and a firm headquartered in these types of MSAs. In Column 2, the estimated coefficient of the *LogClimate* \times *LogAge* is on average -0.628 and is again 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 \(2005\)](#)).²⁸

In Column 3, the average estimated coefficient of the *LogRecreation* \times *LogAge* is -0.865 , with an average t -statistic of -4.03 . The first-stage F -statistic is on average 16.42 . Hence, this instrument is a bit stronger than the previous one. In Column 4, the first-stage F -statistic also increases to 21.37 as we include the other instruments. Our two best instruments remain strong in the presence of additional excluded interactions.

In Panel B, we examine the change of the *Distance* coefficient estimate in the linear under-diversification model of Equation (9), when we switch from OLS to 2SLS. In Column 1, where the endogeneity of *Distance* is not corrected, the OLS coefficient equals on average -0.101 and has a t -statistic of -5.73 . It is a well-established result in the literature that distance emerges as one of the most significant explanatory variables for under-diversification. The estimated coefficients of the control variables are very similar to the ones in previous studies (e.g., [Goetzmann and Kumar \(2008\)](#)), where the linear under-diversification setup has been applied.

In Column 2, where the instrument is only the interaction *LogClimate* \times *LogAge*, the

²⁸The 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.

just-identified estimate of the *Distance* coefficient is on average -0.082 . This is a 19% decrease relative to the uncorrected case. The average t -statistic is -1.92 . In Column 3, where the instrument is only the interaction $\text{LogRecreation} \times \text{LogAge}$, the just-identified estimate of the *Distance* coefficient is on average -0.077 and has an average t -statistic of -1.84 . The reduction relative to the uncorrected case is now 24%.²⁹

In Column 4, we present the estimation results from 2SLS regressions that use as instruments all the interactions between the climate and recreation scores in the MSAs of the stocks' headquarters and the Household Basic Demographics (i.e., age, marital status, number of children, and gender). The obtained estimate of the *Distance* coefficient becomes on average -0.074 , which corresponds to a 27% reduction relative to the uncorrected OLS estimate. The average t -statistic is -1.90 .

Panel B also contains the p -value of a Hansen J -test for overidentifying restrictions. On average, it equals 0.71. Thus, the null hypothesis that the instruments are exogenous is not rejected at any reasonable level of statistical significance.

In Table 6, we compare the average coefficient estimates and t -statistics of 2SLS (in Column 2) with the ones obtained by LIML (in Column 3). Both estimation methods yield virtually identical results. The coefficient estimate of *Distance* is on average -0.073 and has an average t -statistic of -1.91 . For robustness, we close this subsection with the GMM estimation of the linear portfolio under-diversification model. The results are tabulated in Column 4 of Table 6. 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.

In Subfigure 1b, we present the complete time series evolution of the distance coefficient estimates (multiplied by 100), 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

²⁹The 5% additional reduction in the *Distance* coefficient can be attributed to $\text{LogRecreation} \times \text{LogAge}$ being a stronger instrument than $\text{LogClimate} \times \text{LogAge}$, as the F -statistics in Panel A of Table 5 indicate.

when the instrument is $\text{LogClimate} \times \text{LogAge}$, and a green line when the instrument is $\text{LogRecreation} \times \text{LogAge}$. With few exceptions (e.g., first quarter of 1991, second quarter of 1992, etc.), both IV estimates lie quite close to each other. The GMM coefficient estimates of local bias over time are depicted with a yellow line and lie, almost always, below the red and green lines of the just-identified IV estimates.

We also check the results from the reduced-form estimation, where *Distance* is replaced with each one of the excluded household-MSA interactions in the portfolio under-diversification model. In Column 2, the average estimated coefficient of $\text{LogRecreation} \times \text{LogAge}$ is 0.056, while in Column 3, the average estimated coefficient of $\text{LogRecreation} \times \text{LogAge}$ is 0.07. The average *t*-statistics in both cases are higher than 2, i.e. 2.51 and 2.85 respectively, so that the coefficients of these interactions are statistically significant. Moreover, we find that the two key interactions have a similar magnitude in a reduced form in which *Distance* is replaced by *all* the instruments — as in Column 3. Their average *t*-statistics become actually slightly higher.

4.4. Further Analysis on Exclusion Restriction

As we mentioned earlier, our exclusion is a plausible one judged by using traditional household portfolio choice models. However, there are non-traditional models where households select stocks based on the advertisement that they see in magazines which could conceivably violate our exclusion restriction. As a robustness exercise, in Table 7, we present the results that we obtain when we re-estimate the linear and non-linear model in subsamples of stocks excluding consumer-oriented companies. Specifically, in Columns 1 and 2, we drop from the analysis stocks in the industries of food, consumption, cars, retail stores and services. On the other hand, in Columns 3 and 4, we drop stocks of companies that have high advertising expenses. These are firms whose advertising budget is higher than or equal to the average advertising expenses in a given year as well as firms which offer sports sponsorships in the

major league sports.³⁰

In Panel A, we see that the corrections in the Tobit model are 44% and 50% in the first and second subsample of stocks respectively. These numbers are similar to the full sample results. Panel B shows that the correction in the distance coefficient in the linear model is 30% when we exclude the stocks of consumer-oriented companies, and 40% when we exclude the stocks of the big advertisers.

5. Location Choice Residuals and Household Portfolios

Our analysis points to location-choice model residuals being natural proxies for these latent or unobserved economic expectations implicit in location choices. Such latent expectations data would be valuable and hard to find. We can infer them from location choices that are unexplained from observables. These latent expectations more precisely can be captured by Pearson residuals from location-choice models. We show that these residuals are not only natural but powerful proxies empirically, by estimating a structural model of portfolio choice that 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 then estimate this model using the residuals from the location choice model as proxies for these priors.

5.1. Model

Specifically, following [Hong and Xu \(2018\)](#), 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, \mathbf{X}_j , and his expectation about

³⁰The annual advertising expenses of the firms are obtained from Compustat and Advertising Age, while the list of publicly-traded sports sponsors is from [Branikas \(2018\)](#).

stock j 's headquarters' city, $L_{i,j}$:

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

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

$$S_{i,j} = \tilde{f}_j + \eta_{i,j} \quad (21)$$

where $\eta_{i,j} | \mathbf{X}_j, L_{i,j}, dist_{i,j} \sim \mathcal{N} (0, 1/\tau^\eta (dist_{i,j}))$. That is, conditional on stock j 's risk factors, \mathbf{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:

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

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

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

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:

$$\xi_{i,j} | \mathbf{X}_j, L_{i,j}, dist_{i,j} \sim \mathcal{N} \left(0, \underbrace{\frac{\tau^\eta (dist_{i,j})}{\tau^0 + \tau^\eta (dist_{i,j})}}_{\sigma_{i,j}^2} \right) \quad (24)$$

so that the variance of investor i 's latent demand for stock j (namely, $\sigma_{i,j}^2$) 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).³¹

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

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

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:

$$\Lambda_i = \frac{c}{W_i^2} \quad (26)$$

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:³²

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

The KKT conditions then imply that:

³¹That is, $\frac{d\sigma_{i,j}^2(dist_{i,j})}{ddist_{i,j}} = \frac{\tau^n(dist_{i,j})'}{[\tau^0 + \tau^n(dist_{i,j})]^2} < 0$, if $\tau^n(dist_{i,j})' < 0$.

³²Investor 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.

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

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 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:

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

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

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

where $\zeta_{i,j} | \mathbf{X}_j, L_{i,j}, dist_{i,j} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. 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 3.1. That is, we define:

$$L_{i,j} \equiv \frac{r_{i,j} - p_{i,j}}{\sqrt{p_{i,j}}} \quad (31)$$

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. Structural 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 8, 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 15.55.³³

Panel B of Table 8 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

³³The 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.

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 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 on latent subjective expectations about the economic prospects of a city, which are 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 a natural and powerful proxy for such latent or subjective expectations using Pearson residuals from location-choice models.

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. We believe the use of these Pearson residuals from location-choice models can be an important part of household finance toolkit. 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.

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.
- Akaike, H., 1974, “A New Look at the Statistical Model Identification,” *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Angrist, J. D., and J.-S. Pischke, 2008, *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton University Press.
- 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. 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.
- 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.
- Branikas, I., 2018, “Advertising Exposure and Portfolio Choice: Estimates Based on Sports Sponsorships,” *Working Paper*.

- 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.
- 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.
- 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.
- Hong, H., and J. Xu, 2018, “Inferring Latent Social Networks from Stock Holdings,” *Journal of Financial Economics*.
- 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, 2017, “Understanding the Long-Run Decline in Interstate Migration,” *International Economic Review*, 58(1), 57–94.
- Keloharju, M., S. Knupfer, and E. Rantapuska, 2012, “Mutual Fund and Share Ownership in Finland,” *Liiketaloudellinen aikakauskirja*, 2, 178–198.

- Korniotis, G. M., and A. Kumar, 2011, “Do Older Investors Make Better Investment Decisions?,” *Review of Economics and Statistics*, 93(1), 244–265.
- , 2013, “Do Portfolio Distortions Reflect Superior Information or Psychological Biases?,” *Journal of Financial and Quantitative Analysis*, 48(1), 1–45.
- 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.
- McFadden, D., 1978, “Modeling the Choice of Residential Location,” *Transportation Research Record*, (673).
- 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.
- Sinha, P., M. L. Caulkins, and M. L. Cropper, 2017, “Household Location Decisions and the Value of Climate Amenities,” *Journal of Environmental Economics and Management*.
- Stock, J., and M. Yogo, 2005, “Testing for Weak Instruments in Linear IV Regression,” in *Identification and Inference for Econometric Models*, ed. by D. W. Andrews. Cambridge University Press, New York, pp. 80–108.
- 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 summarizes all the variables in our sample. 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. The "Household Basic Demographics" are: 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. The "Household Prof. Demographics" are: Income, i.e. the income of the household's head, Education, i.e. the percentage of the household's ZIP-Code population that holds a bachelor's or higher degree, ProfProxim, i.e. the professional industrial proximity to a stock of the household's ZIP-Code population, and the occupation codes of the household's head, i.e. Professional, Managerial, SalesSvc, WhiteCollar and BlueCollar — which are indicator variables 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. Panel C 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. The stock industries are defined based on the 17 Fama-French industry portfolios. Food to SvcOth are indicator variables that equal one if a stock belongs to the corresponding industry. Panel D 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 E refers to the local bias of household portfolio holdings. Column 1 (Column 2) reports the average distance of households from the stocks in their portfolios (in a benchmark). In Row 1 (Row 2), the benchmark is the equally weighted (value-weighted) portfolio. Column 3 reports the difference of Column 2 from Column 1, i.e. the local bias in distance units. Column 4 reports the local bias (LB) measure as a percentage. Column 5 reports the t -statistics for the LB measure. 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
Unemp (%)	6.25	1.71	5.97	2.92	15.63
HPI	94.8	10.63	94.32	66.04	122.63
Pop (million)	2.46	2.71	1.52	0.68	17.68
Transportation	4,816.47	1,235.59	4,705	6.97	7,429
Colleges	1,523.38	1,400.53	1,122	1.45	6,728
HealthCare	298.14	140.27	274	109	880
Crime	1,219.02	399.44	1,165	550	2,821
Recreation	2,130.49	787.78	2,104	707	3,940
Climate	577.02	116.71	559	287	910

Table Cont'd: Summary Statistics

Panel B: Household Demographics					
	Mean	S.D.	Median	Min	Max
<u>HH Basic Demo's</u>					
Age	51.78	12.51	50	21	80
FamSize	2.49	1.2	2	1	6
Male	0.92	0.28	1	0	1
Married	0.73	0.44	1	0	1
<u>HH Prof. Demo's</u>					
Income (thousand \$)	101.85	68.24	87.5	10	250
Education	0.36	0.15	0.35	0	0.91
ProfProxim	0.08	0.06	0.07	0	0.56
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
<u>HH ZIP-Code Race Pct.</u>					
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
Panel C: Stock Financial Characteristics					
	Mean	S.D.	Median	Min	Max
Size (million \$)	3589.48	7962.3	1245.65	1.7	159758.9
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
SvcOth	0.25	0.44	0	0	1

Table Cont'd: Summary Statistics

Panel D: 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 E: Local Bias among Households					
	Avg. Distance from				
Weights	(1) Holdings	(2) Benchmark	(3) Difference	(4) % Bias (LB)	(5) <i>t</i> -stat
Equal	9.38	17.67	8.29	45.45	52.74
Value	9.38	17.65	8.26	43.72	47.65

Table 2: Balance Tests Based on Climate and Recreation

This table presents the balance-tests based on recreation and climate. In Panel A (B), the MSA sample is split into two groups based on the median climate (recreation) score. In each subsample, the averages of the MSA demographics are calculated. 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 tests for these variables in every year and present the time-series average means, differences and t -statistics. The sample consists of 57 MSAs with a population of at least 750K at the end of 1996. Panel C presents the estimation results of two-variable regressions of stock financial characteristics on the log of recreation (Columns 1 and 2) and climate score (Columns 3 and 4) 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).

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
Panel C: Bivariate Regressions				
Depend. Variable	Independ. Variable: LogClimate		Independ. Variable: LogRecreation	
	(1) Coef. Est.	(2) t -statistic	(3) Coef. Est.	(4) 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
SvcOth	0.11	1.7	0.058	1.62

Table 3: Conditional Logistic Regressions of Household Location Choice

This table presents the conditional logistic regressions of household location choice. The dependent variable is $r_{i,c}$, i.e. an indicator variable that equals one if household i resides in MSA c . The independent variables are $\text{LogClimate} \times \text{LogAge}$, $\text{LogRecreation} \times \text{LogAge}$ and the MSA climate and recreation scores. Other MSA Demographics and their interactions with the Household Basic Demographics are included as controls. See Table 1 for a detailed description. In Column 3, the interactions of the (log) climate and recreation scores of the MSAs with the other Household Basic Demographics (i.e., LogFamSize , Male and Married) are also included. A likelihood ratio test is performed in Columns 1 and 2 for restricted versions of the full model in Column 3. AIC is the Akaike information criterion. The table depicts the coefficient estimates and t -statistics [in brackets] based on standard errors clustered at the household level.

	(1)	(2)	(3)
$\text{LogClimate} \times \text{LogAge}$	0.459 [4.02]		0.436 [4.39]
$\text{LogRecreation} \times \text{LogAge}$		0.633 [4.67]	0.571 [4.28]
LogClimate	2.373 [2.41]	0.775 [11.31]	3.243 [2.59]
LogRecreation	0.281 [7.27]	-0.629 [-1.15]	-1.772 [-2.6]
Other MSA Demo's	YES	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES	YES
$\text{LogClimate} \times \{\text{LogFamSize}, \text{Male}, \text{Married}\}$	NO	NO	YES
$\text{LogRecreation} \times \{\text{LogFamSize}, \text{Male}, \text{Married}\}$	NO	NO	YES
LR Test vs. Full Model	126.97	75.15	-
AIC	61,883	61,831	61,770
Number of HH	8,688	8,688	8,688
Number of MSAs	57	57	57

Table 4: Tobit Regressions of Household Portfolio Choice Without and With Correction for Location Choice

This table presents Tobit regressions of household portfolio weights. The dependent variable is $w_{i,c,h,j}$, i.e. the portfolio weight of household i residing in city c on stock j headquartered in city h . The independent variable is $Distance$, i.e. the distance (in degrees) of stock j 's headquarters' ZIP-Code from household i 's address ZIP-Code. The controls include stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of the Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. In Column 1, there are no control functions, so location choice is not taken into account. In Column 2, the approximation of the control functions is cubic. In Column 3, the approximation of the control functions is quartic. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	(1)	(2)	(3)
	Uncor.	Cubic Ψ 's	Quartic Ψ 's
Distance	-0.014 [-6.13]	-0.008 [-11.33]	-0.008 [-14.12]
LogSize	0.364 [47.32]	0.361 [53.08]	0.36 [58.04]
BTM	0.05 [6.27]	0.051 [5.73]	0.051 [5.65]
Turnover	0.541 [11.28]	0.524 [9.16]	0.51 [8.66]
Momentum	-0.198 [-8.99]	-0.196 [-8.3]	-0.197 [-8.14]
Volatility	2.861 [18.43]	2.836 [17.39]	2.815 [17.33]
Profitability	-0.066 [-1.25]	-0.071 [-1.5]	-0.076 [-1.13]
Investment	-0.097 [-3.96]	-0.097 [-5.85]	-0.099 [-5.94]
Stock Industry FE	YES	YES	YES
Household Basic Demo's	YES	YES	YES
Household Prof. Demo's	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES
LogClimate	YES	YES	YES
LogRecreation	YES	YES	YES
Other MSA Demo's	YES	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES	YES
Average Number of Households	4,339	4,339	4,339
Average Number of Stocks	988	988	988

Table 5: OLS vs. 2SLS Regressions of Household Excess Portfolio Weight on Distance

This table presents the OLS versus 2SLS regressions of household excess portfolio weights. Panel A depicts the first-stage. The dependent variable is *Distance*, i.e. the distance (in degrees) of stock *j*'s headquarters' ZIP Code from household *i*'s address. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The controls include stock financial characteristics, household demographics, demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores — and interactions of the Household Basic Demographics with all *other* MSA demographics. See Table 1 for a detailed description. Column 1 shows the OLS. In Column 2, the instrument is $\text{LogClimate} \times \text{LogAge}$. In Column 3, the instrument is $\text{LogRecreation} \times \text{LogAge}$. In Column 4, the instruments are *all* the interactions of the (log) climate and recreation scores of the MSAs of the stocks' headquarters with the Household Basic Demographics. The first-stage *F*-statistic tests the hypothesis that the coefficient(s) of the instrument(s) is (are jointly) zero — taking into account the two-level clustering. Panel B depicts the second-stage. The dependent variable is $EW_{i,c,h,j} \equiv (w_{i,c,h,j} - w_j^{VW})/w_j^{VW}$, i.e. the excess portfolio weight of household *i* residing in city *c* on stock *j* headquartered in city *h* (relative to the market value-weighted portfolio weight on stock *j*). The Hansen *J*-test is the test of overidentifying restrictions. The estimation is performed in a panel of households and stocks for every month separately. The table depicts the average monthly coefficient estimates and the average *t*-statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	(1)	(2)	(3)	(4)
Panel A: First-Stage				
LogClimate \times LogAge		-0.628 [-3.65]		-0.516 [-3.99]
LogRecreation \times LogAge			-0.865 [-4.03]	-0.75 [-3.82]
Stock Financial Char's		YES	YES	YES
Household Basic Demo's		YES	YES	YES
Household Prof. Demo's		YES	YES	YES
HH ZIP-Code Race Pct.		YES	YES	YES
LogClimate		YES	YES	YES
LogRecreation		YES	YES	YES
Other MSA Demo's		YES	YES	YES
Other MSA Demo's \times HH Basic Demo's		YES	YES	YES
Other Instruments		NO	NO	YES
<i>F</i> -statistic		13.45	16.42	21.37
Average Number of Households	4,339	4,339	4,339	4,339
Average Number of Stocks	988	988	988	988
Panel B: Second-Stage				
	OLS	2SLS	2SLS	2SLS
Distance	-0.101 [-5.73]	-0.082 [-1.92]	-0.077 [-1.84]	-0.074 [-1.90]
Stock Financial Char's	YES	YES	YES	YES
Household Basic Demo's	YES	YES	YES	YES
Household Prof. Demo's	YES	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES	YES
LogClimate	YES	YES	YES	YES
LogRecreation	YES	YES	YES	YES
Other MSA Demo's	YES	YES	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES	YES	YES
Hansen <i>J</i> -test (<i>p</i> -value)				0.71
Average Number of Households	4,339	4,339	4,339	4,339
Average Number of Stocks	988	988	988	988

Table 6: OLS vs. 2SLS, LIML & GMM Regressions of Household Excess Portfolio Weights on Distance

This table presents the OLS vs. 2SLS, LIML and GMM regressions of household excess portfolio weights. The dependent variable is $EW_{i,c,h,j} \equiv (w_{i,c,h,j} - w_j^{YW})/w_j^{YW}$, i.e. the excess portfolio weight of household i residing in city c on stock j headquartered in city h (relative to the market value-weighted portfolio weight on stock j). The independent variable is *Distance*, i.e. the distance (in degrees) of stock j 's headquarters' ZIP-Code from household i 's address ZIP-Code. The controls are stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of the Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. Column 1 shows the OLS, Column 2 the 2SLS, Column 3 the LIML, and Column 4 the GMM. 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 [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	(1)	(2)	(3)	(4)
	OLS	2SLS	LIML	GMM
Distance	-0.101 [-5.73]	-0.074 [-1.90]	-0.073 [-1.91]	-0.067 [-1.95]
LogSize	-0.608 [-5.82]	-0.606 [-6.23]	-0.606 [-6.23]	-0.544 [-6.74]
BTM	-0.367 [-0.32]	-0.368 [-0.32]	-0.368 [-0.32]	-0.186 [-0.19]
Turnover	2.669 [2.21]	2.202 [1.81]	2.202 [1.81]	2.962 [2.71]
Momentum	-1.557 [-5.64]	-1.543 [-6.01]	-1.543 [-6.01]	-1.407 [-6.24]
Volatility	29.863 [5.38]	29.342 [5.54]	29.342 [5.54]	24.008 [5.76]
Profitability	-0.494 [-0.93]	-0.58 [-1.5]	-0.58 [-1.5]	-0.425 [-1.28]
Investment	-0.526 [-3.34]	-0.53 [-3.5]	-0.53 [-3.5]	-0.445 [-3.59]
Stock Industry FE	YES	YES	YES	YES
Household Basic Demo's	YES	YES	YES	YES
Household Prof. Demo's	YES	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES	YES
LogClimate	YES	YES	YES	YES
LogRecreation	YES	YES	YES	YES
Other MSA Demo's	YES	YES	YES	YES
Other MSA Demo's × HH Basic Demo's	YES	YES	YES	YES
Average Number of Households	4,339	4,339	4,339	4,339
Average Number of Stocks	988	988	988	988

Table 7: Regressions of Household Portfolio Weights on Distance In Selected Subsamples of Stocks

This table presents the Tobit regressions of household portfolio weights and the linear regressions of household excess portfolio weights in selected subsamples of stocks. In Columns 1 and 2, we drop the stocks of consumer-oriented companies, and in Columns 3 and 4 the stocks of companies with high advertising expenditure. In Panel A, the dependent variable is $w_{i,c,h,j}$, i.e. the portfolio weight of household i residing in city c on stock j headquartered in city h . The independent variable is *Distance*, i.e. the distance of stock j 's headquarters' ZIP Code from household i 's address ZIP Code. The controls include stock financial characteristics, household demographics and demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of the Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. Columns 1 and 3 do not correct for household location choice. Columns 2 and 4 correct for household location choice, approximating the control functions with fourth order polynomials. In Panel B, the dependent variable is $EW_{i,c,h,j} \equiv (w_{i,c,h,j} - w_j^{VW})/w_j^{VW}$, i.e. the excess portfolio weight of household i residing in city c on stock j headquartered in city h (w.r.t. to the market value-weighted portfolio weight on stock j). Columns 1 and 3 show the OLS, and Columns 2 and 4 the GMM. 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 estimation is performed in a panel of households and stocks for every month separately. The table depicts the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	No Consumer-Oriented Stocks		No Stocks with High AD	
	(1)	(2)	(3)	(4)
Panel A: Non-linear Portfolio Choice Model				
	Uncor.	Corrected	Uncor.	Corrected
Distance	-0.016 [-3.67]	-0.009 [-5.65]	-0.022 [-4.92]	-0.011 [-7.22]
Stock Financial Char's	YES	YES	YES	YES
Household Basic Demo's	YES	YES	YES	YES
Household Prof. Demo's	YES	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES	YES
LogClimate	YES	YES	YES	YES
LogRecreation	YES	YES	YES	YES
Other MSA Demo's	YES	YES	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES	YES	YES
Average Number of Households	4,339	4,339	4,339	4,339
Average Number of Stocks	988	988	988	988
Panel B: Linear Portfolio Under-Diversification Model				
	OLS	GMM	OLS	GMM
Distance	-0.087 [-5.69]	-0.061 [-2.12]	-0.091 [-6.79]	-0.055 [-2.26]
Stock Financial Char's	YES	YES	YES	YES
Household Basic Demo's	YES	YES	YES	YES
Household Prof. Demo's	YES	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES	YES
LogClimate	YES	YES	YES	YES
LogRecreation	YES	YES	YES	YES
Other MSA Demo's	YES	YES	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES	YES	YES
First-stage F -statistic		17.3		19.1
Hansen J -test (p -value)		0.82		0.75
Average Number of Households	4,339	4,339	4,339	4,339
Average Number of Stocks	988	988	988	988

Table 8: Structural Estimation of Retail Investor Portfolio Choice

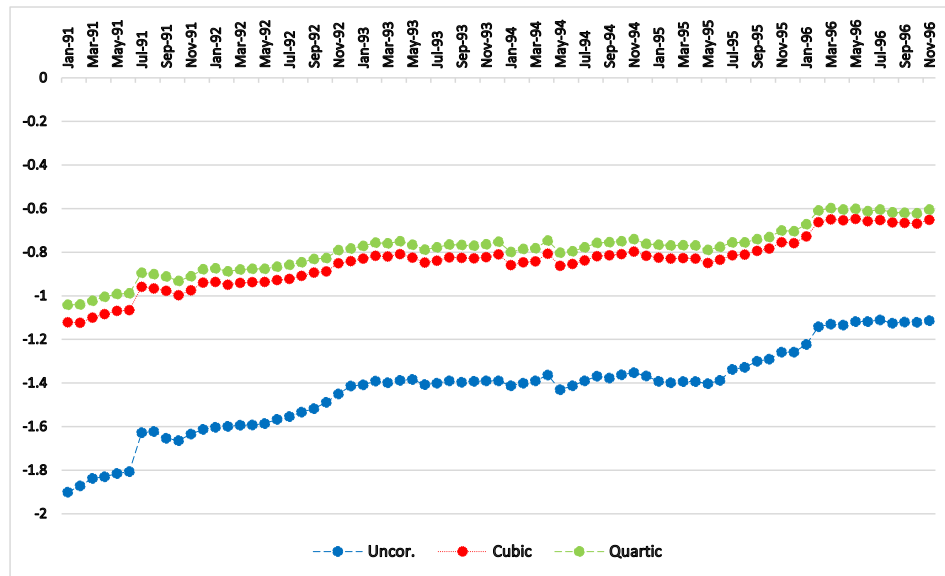
This table presents the structural estimation of retail investor portfolio choice. The dependent variable is $w_{i,j}$, i.e. the portfolio weight of household i on stock j . The independent variables are *Distance*, i.e. the distance between household i 's residential ZIP-Code and the ZIP-Code of stock j 's headquarters, and *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 estimation is performed in a panel of households and stocks for every month separately. Panel A shows the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household. A likelihood ratio test is performed in Columns 2 and 3 for restricted versions of the full model in Column 1. Panel B shows the average economic effect of *Distance* and *LocRes* on $w_{i,j}$ in basis points and as percentage of the mean.

	(1)	(2)	(3)
	Full Model	No LocRes	No Distance
Panel A: Coefficient Estimates			
Distance	-0.002 [-3.04]	-0.004 [-5.65]	
LocRes	0.126 [16.75]		0.138 [15.55]
LogSize	0.372 [48.13]	0.373 [44.65]	0.373 [47.59]
BTM	0.052 [6.37]	0.054 [6.38]	0.052 [6.08]
Turnover	0.65 [13.06]	0.689 [13.25]	0.64 [11.84]
Momentum	-0.191 [-9.44]	-0.192 [-8.7]	-0.192 [-9.23]
Volatility	3.065 [23.03]	3.1 [21.64]	3.054 [21.88]
Profitability	-0.029 [-0.62]	-0.012 [-0.17]	-0.033 [-0.72]
Investment	-0.082 [-4.68]	-0.079 [-4.26]	-0.084 [-4.76]
Stock Industry FE	YES	YES	YES
LR-test vs. Full Model		2,185.5	212.48
Average Number of Households	4,339	4,339	4,339
Average Number of Stocks	988	988	988
Panel B: Economic Effects in bps and as % of the Mean			
Distance	-0.98 -9.5%	-2.25 -21.8%	
LocRes	1.5 14.5%		1.68 16.3%

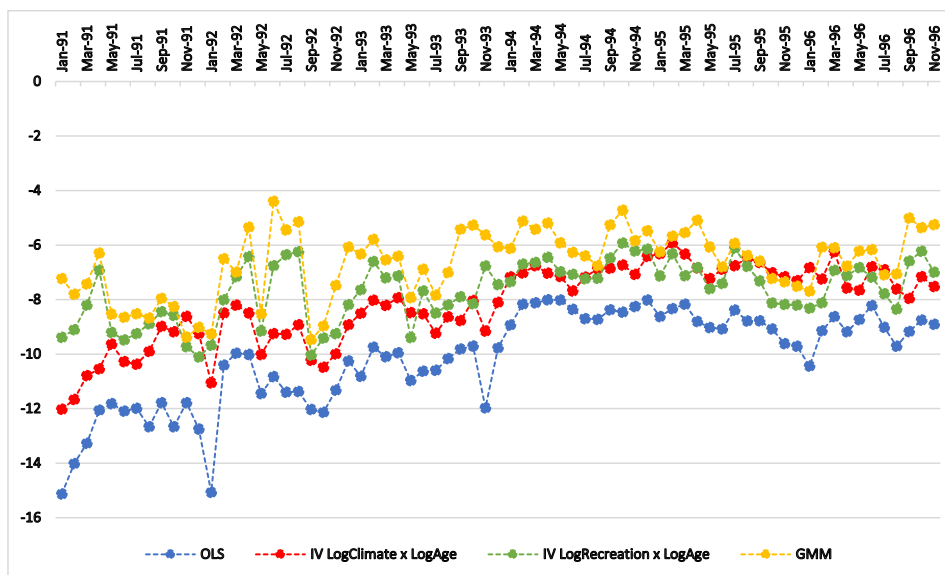
Figure 1: Coefficient Estimates of Distance Over Time in Non-Linear Regressions and Linear of Household Portfolio Weights

The figures depict the coefficient estimates of *Distance* multiplied by 100 — on the vertical axis — across the time periods — on the horizontal axis — in the non-linear and linear regressions of portfolio weights with full controls. Subfigure 1 refers to three Tobit regressions of portfolio weights ($w_{i,c,h,j}$). The blue line depicts the estimates without correcting for location choice. The red line depicts the estimates with a third order approximation of the correction functions. The green line depicts the estimates with a fourth order approximation of the correction functions. Subfigure 2 refers to four linear regressions of household excess portfolio weights ($EW_{i,c,h,j}$). The blue line depicts the OLS estimates without correcting for location choice. The red line depicts the just-identified estimate from the instrument $LogClimate \times LogAge$. The green line depicts the just-identified estimate from the instrument $LogRecreation \times LogAge$. The yellow line depicts the GMM estimates with instruments the interactions of $LogClimate$ and $LogRecreation$ with all the household location demographics.

(a) Distance Coef. in the Non-Linear Portfolio Choice Model



(b) Distance Coef. in the Linear Under-Diversification Model



Online 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.

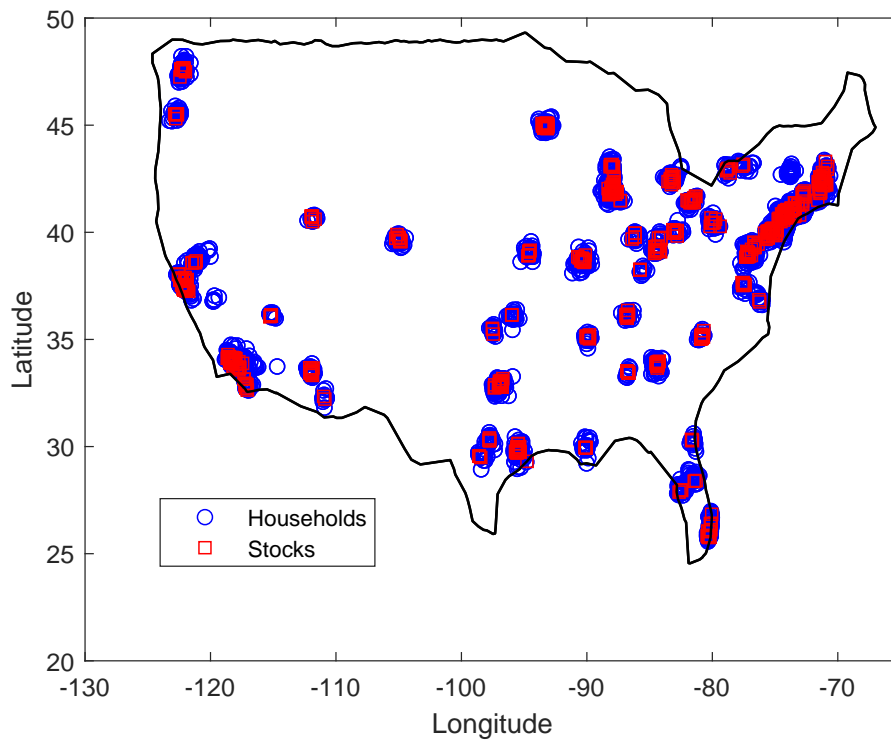
Crime: This score is based on the violent crime rate and the property crime rate divided by 10. The lower the score of crime, the less the crime in the MSA.

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.

Appendix 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.



Robustness Checks and Extensions

In this Online 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 reduction 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 estimation 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 -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 Tobit regressions of household portfolio weights on distance indicator variables. The dependent variable is $w_{i,c,h,j}$, i.e. the portfolio weight of household i residing in city c on stock j headquartered in city h . The independent 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 and 2, the distance threshold is 100 miles. In Columns 3 and 4, the distance threshold is 250 miles. The controls include stock financial characteristics, household demographics, demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. Columns 1 and 3 do not correct for household location choice. Columns 2 and 4 correct for household location choice, approximating the control functions with fourth order polynomials. The estimation is performed in a panel of households and stocks for every month separately. The table presents the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	100 Miles Away		250 Miles Away	
	(1)	(2)	(3)	(4)
	Uncor.	Corrected	Uncor.	Corrected
Away	-0.629 [-5.49]	-0.428 [-6.73]	-0.504 [-5.04]	-0.295 [-7.35]
LogSize	0.362 [46]	0.36 [47.07]	0.363 [49.16]	0.36 [50.76]
BTM	0.047 [3.34]	0.05 [2.96]	0.051 [3.65]	0.052 [4.08]
Turnover	0.5 [11.12]	0.494 [8.26]	0.525 [10.24]	0.506 [8.64]
Momentum	-0.197 [-5.62]	-0.196 [-5.92]	-0.196 [-5.68]	-0.196 [-5.92]
Volatility	2.847 [4.29]	2.807 [5.22]	2.821 [3.91]	2.795 [5.79]
Profitability	-0.079 [-7.45]	-0.081 [-8.4]	-0.07 [-4.14]	-0.076 [-5.54]
Investment	-0.099 [-4.12]	-0.1 [-4.3]	-0.099 [-3.73]	-0.1 [-4.85]
Stock Industry FE	YES	YES	YES	YES
HH Basic Demo's	YES	YES	YES	YES
HH Prof. Demo's	YES	YES	YES	YES
HH Occupation-Code FE	YES	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES	YES
LogClimate	YES	YES	YES	YES
LogRecreation	YES	YES	YES	YES
Other MSA Demo's	YES	YES	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES	YES	YES
Average Number of Households	4,339	4,339	4,339	4,339
Average Number of Stocks	988	988	988	988

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

This table presents Tobit regressions of household portfolio weights on the natural logarithm of distance. The dependent variable is $w_{i,c,h,j}$, i.e. the portfolio weight of household i residing in city c on stock j headquartered in city h . The independent variable is $LogDist$, i.e. the log of the distance of stock j 's headquarters' ZIP Code from household i 's address ZIP Code. The controls include stock financial characteristics, household demographics, demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. Columns 1 does not correct for household location choice. Columns 2 corrects for household location choice, approximating the control functions with fourth order polynomials. The estimation is performed in a panel of households and stocks for every month separately. The table presents the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	(1)	(2)
	Uncor.	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]
Stock Industry FE	YES	YES
HH Basic Demo's	YES	YES
HH Prof. Demo's	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct.	YES	YES
LogClimate	YES	YES
LogRecreation	YES	YES
Other MSA Demo's	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES
Average Number of Households	4,339	4,339
Average Number of Stocks	988	988

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

This table presents the estimation results from linear regressions of household excess portfolio weights on distance indicator variables. The dependent variable is $EW_{i,c,h,j} \equiv (w_{i,c,h,j} - w_j^{VW})/w_j^{VW}$, i.e. the excess portfolio weight of household i residing in city c on stock j headquartered in city h (w.r.t. to the market value-weighted portfolio weight on stock j). The independent 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. The controls include stock financial characteristics, household demographics, demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. Columns 1 and 3 shows the OLS, and Columns 2 and 4 the GMM. 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 estimation is performed in a panel of households and stocks for every month separately. The table depicts the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	100 Miles Away		250 Miles Away	
	(1)	(2)	(3)	(4)
	OLS	GMM	OLS	GMM
Away	-7.525 [-3.98]	-5.389 [-1.7]	-4.86 [-4.84]	-3.328 [-1.82]
LogSize	-0.613 [-6.08]	-0.545 [-6.71]	-0.611 [-6.02]	-0.548 [-6.81]
BTM	-0.396 [-0.38]	-0.109 [-0.18]	-0.377 [-0.34]	-0.12 [-0.14]
Turnover	1.857 [1.56]	2.895 [2.46]	2.116 [1.78]	2.861 [2.65]
Momentum	-1.536 [-5.64]	-1.403 [-6.19]	-1.541 [-5.65]	-1.398 [-6.21]
Volatility	29.333 [5.24]	24.301 [5.8]	29.257 [5.24]	24.15 [5.79]
Profitability	-0.621 [-1.24]	-0.427 [-1.26]	-0.549 [-1.06]	-0.407 [-1.2]
Investment	-0.539 [-3.51]	-0.461 [-3.7]	-0.541 [-3.51]	-0.459 [-3.74]
Stock Industry FE	YES	YES	YES	YES
HH Basic Demo's	YES	YES	YES	YES
HH Prof. Demo's	YES	YES	YES	YES
HH Occupation-Code FE	YES	YES	YES	YES
HH ZIP-Code Race Pct.	YES	YES	YES	YES
LogClimate	YES	YES	YES	YES
LogRecreation	YES	YES	YES	YES
Other MSA Demo's	YES	YES	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES	YES	YES
First-stage F -statistic		25.42		23.01
Hansen J -test (p -value)		0.69		0.73
Average Number of Households	4,339	4,339	4,339	4,339
Average Number of Stocks	988	988	988	988

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

This table presents linear regressions of household excess portfolio weights on the natural logarithm of distance. The dependent variable is $EW_{i,c,h,j} \equiv (w_{i,c,h,j} - w_j^{VW})/w_j^{VW}$, i.e. the excess portfolio weight of household i residing in city c on stock j headquartered in city h (w.r.t. to the market value-weighted portfolio weight on stock j). The independent variable is $LogDist$, i.e. the log of the distance of stock j 's headquarters' ZIP Code from household i 's address ZIP Code. The controls include stock financial characteristics, household demographics, demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. Column 1 shows the OLS and Column 2 shows the GMM. 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 estimation is performed in a panel of households and stocks for every month separately. The table depicts the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	(1)	(2)
	OLS	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]
Stock Industry FE	YES	YES
HH Basic Demo's	YES	YES
HH Prof. Demo's	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct.	YES	YES
LogClimate	YES	YES
LogRecreation	YES	YES
Other MSA Demo's	YES	YES
Other MSA \times HH Basic Demo's	YES	YES
First-stage F -statistic		26.41
Hansen J -test (p -value)		0.58
Average Number of Households	4,339	4,339
Average Number of Stocks	988	988

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

This table presents two Tobit regressions of household portfolio weights. 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. The dependent variable is $w_{i,c,h,j}$, i.e. the portfolio weight of household i residing in city c on stock j headquartered in city h . The independent variable is *Distance*, i.e. the distance of stock j 's headquarters' ZIP Code from household i 's address ZIP Code. The controls include stock financial characteristics, household demographics, demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. Column 1 does not correct for household location choice. Column 2 corrects for household location choice, approximating the control functions with fourth order polynomials. The estimation is performed in a panel of households and stocks for every month separately. The table presents the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	(1)	(2)
	Uncor.	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]
Stock Industry FE	YES	YES
HH Basic Demo's	YES	YES
HH Prof. Demo's	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct.	YES	YES
LogClimate	YES	YES
LogRecreation	YES	YES
Other MSA Demo's	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES
Average Number of Households	5,524	5,524
Average Number of Stocks	3,517	3,517

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

This table presents linear regressions of household excess portfolio weights. 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. The dependent variable is $EW_{i,c,h,j} \equiv (w_{i,c,h,j} - w_j^{VW})/w_j^{VW}$, i.e. the excess portfolio weight of household i residing in city c on stock j headquartered in city h (w.r.t. to the market value-weighted portfolio weight on stock j). The independent variable is $Distance$, i.e. the distance of stock j 's headquarters' ZIP Code from household i 's address ZIP Code. The controls include stock financial characteristics, household demographics, demographics of the MSAs of the stocks' headquarters — including the (log) climate and recreation scores. The instruments are pair-wise interactions between the (log) climate and recreation scores of the MSAs of the stocks' headquarters and the Household Basic Demographics. The interactions of Household Basic Demographics with all *other* MSA demographics are included as controls. See Table 1 for a detailed description. Column 1 shows the OLS and Column 2 the GMM. 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 estimation is performed in a panel of households and stocks for every month separately. The table depicts the average monthly coefficient estimates and the average t -statistics [in brackets] based on two-way clustered standard errors at the level of the household and the MSA of the household.

	(1)	(2)
	OLS	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]
Stock Industry FE	YES	YES
HH Basic Demo's	YES	YES
HH Prof. Demo's	YES	YES
HH Occupation-Code FE	YES	YES
HH ZIP-Code Race Pct.	YES	YES
LogClimate	YES	YES
LogRecreation	YES	YES
Other MSA Demo's	YES	YES
Other MSA Demo's \times HH Basic Demo's	YES	YES
First-stage F -statistic		24.39
Hansen J -test (p -value)		0.74
Average Number of Households	5,524	5,524
Average Number of Stocks	3,517	3,517