Superstition and Real Estate Prices: Evidence from a Machine Learning Name Classifier

Brad R. Humphreys *
West Virginia University

Adam Nowak†
West Virginia University

Yang Zhou ‡
West Virginia University

December 10, 2018

Abstract

We investigate the impact of superstition on prices paid by Chinese-American home buyers. Chinese consider *8* lucky and *4* unlucky. Lacking explicit buyer ethnicity identifiers, we develop a binomial name classifier, a machine learning approach applicable to any data set containing names, that allows for falsification tests using other ethnic groups, and mitigates ambiguity from transliteration of Chinese characters into the Latin alphabet. Chinese buyers pay 1-2% premiums for addresses including an *8* and 1% discounts for addresses including a *4*. These results are unrelated to unobserved property quality; no premium exists when Chinese sell to non-Chinese. The persistence of superstitions reflects the extent of cultural assimilation.

JEL Codes: D03, R21, R30
Key words: superstition; supervised learning; name matching; hedonic price model

*John Chambers College of Business & Economics, Department of Economics, 1601 University Ave., PO Box 6025, Morgantown, WV 26506-6025, USA; Email: brhumphreys@mail.wvu.edu

†John Chambers College of Business & Economics, Department of Economics, 1601 University Ave., PO Box 6025, Morgantown, WV 26506-6025, USA; Email: adam.d.nowak@gmail.com

‡Corresponding author. John Chambers College of Business & Economics, Department of Economics, 1601 University Ave., PO Box 6025, Morgantown, WV 26506-6025, USA; Email: ygzhou@mix.wvu.edu

1 Introduction

Anecdotal and empirical evidence suggests that superstitions held by economic agents affect outcomes. For example, less than 5% of condo buildings in New York City contain a 13th floor as *13* is considered an unlucky number in Western Culture.[1] Conversely, a *Lucky Seven Road* can be found in Wisconsin, Pennsylvania, Idaho and Texas. This study investigates preferences for lucky or unlucky numbers specific to Chinese culture — hereafter Chinese *superstition* — in an American real estate market.

We find that individuals with a Chinese cultural background — hereafter *Chinese* — pay a premium (discount) for residential properties with addresses that include lucky (unlucky) numbers in Chinese superstition. The Chinese words for *8* (八 , bā) and *wealth / prosperity* (发 , fā) are phonetically similar. In Chinese superstition, the number *8* is widely believed to be the most lucky single digit. In contrast, the number *4* is considered unlucky as the words for *4* (四, sì) and *death* (死 , sǐ) are also phonetically similar.

Migration, and the assimilation of migrants into their new culture represents an important issue in society (Fernández-Huertas Moraga et al., 2017), and the persistence of cultural superstitions is one indicator of assimilation. It is possible that Chinese who live in America retain their cultural heritage and its associated superstition. Alternatively, some or all Chinese may have completely or partially assimilated into American culture and no longer retain cultural superstition.

We analyze real estate transaction prices based on the digits contained in addresses for single-family homes in the Seattle, Washington metro area. Seattle is an ideal setting for research on cultural assimilation in America as it has been a prime destination for Chinese immigrants since the 1860s and, relative to much of the country, contains a large number of Chinese home buyers and sellers.

In order to determine if Chinese pay a premium or discount for properties based on the presence of specific numbers in addresses, it is first necessary to identify whether or not a buyer or seller is Chinese. Determining ethnicity from names alone represents a major barrier for research on cultural assimilation because many data sources contain names but no information on ethnicity.

We develop a binomial classifier that identifies individuals as Chinese or non-Chinese based on name only. In order to train our classifier, we use a supervised learning algorithm and a labeled data set of Chinese and American participants in the Summer Olympic Games from 1948 to 2012. Intuitively, the binomial classifier is based on the frequency of a given name in the Chinese Olympic rosters relative to the frequency of that name in the US Olympic

[1] Sanette Tanaka, *A 13th Floor Condo? No Such Luck*, Wall Street Journal, September 5, 2013

rosters. Our classification approach uses publicly available data sources. The programs are available at the links below and from the authors upon request.[2] This supervised learning approach to identifying ethnicity is general and can be applied to any data set containing names with no other information about ethnicity.

Because names and genetics are passed on from one generation to the next, procedures for identifying ethnicity have been extensively studied in the biomedical field and are known in general as *name-ethnicity matching* or, when only the surname is used, *surname-ethnicity matching*. In general, the researcher imputes ethnicity using a pre-specified dictionary of names associated with a given ethnicity. Constructing a dictionary of names using frequent names within an ethnic group or frequent names relative to other ethnic groups can be problematic in this setting as many Chinese names are identical to Korean, Vietnamese, and English surnames when Romanized (Quan et al., 2006).[3] For instance, 张 (张, chang) is a common name in China with Wade-Giles Romanization *chang*, while in Korea, 장 is a common name with McCune–Reischauer Romanization *chang*.[4][5] Because of this, out-of-sample mis-classification is possible when the classifier is trained using only two reference groups: Chinese and non-Chinese. Based on this ambiguity in the Romanized names, we develop an alternative multinomial approach than can classify names as either Chinese, US American, or Korean.

Other studies document the relevance of Chinese superstitions in real estate markets, including Chau et al. (2001), Shum et al. (2014), Fortin et al. (2014), Agarwal et al. (2016), and Rehm et al. (2017). Some use data from China (Chau et al., 2001; Shum et al., 2014), or countries with large ethnic Chinese populations like Singapore (Agarwal et al., 2016). Others identify property-level effects using properties located in specific geographic areas with many Chinese residents (Bourassa and Peng, 1999; Fortin et al., 2014) or by simple matching of names to a pre-specified dictionary developed by others (Rehm et al., 2017). We use individual-level data and a machine learning approach to identifying Chinese buyers

[2] A copy of the data and classification program is available from the authors upon request and at
Program: https://dl.dropboxusercontent.com/u/62967289/olympic%20names%20china.R
Auxiliary Program: https://dl.dropboxusercontent.com/u/62967289/fastTDM.R
Olympic Roster Data: https://dl.dropboxusercontent.com/u/62967289/olympicRosters.csv

[3] Romanization refers to the transliteration of non-Latin characters using the Latin alphabet.

[4] Wade-Giles and McCune–Reischauer are Romanization systems for Chinese and Korean characters, respectively.

[5] There have been two main systems of Romanization of Chinese characters in the 20th century. Wade–Giles, the system of transcription in the English-speaking world for most of the 20th century, was developed during the mid-19th century. In 1958, the Pinyin system officially replaced the Wade–Giles system across mainland China and continued to replace Wade-Giles in other Chinese speaking regions of the world. These two systems use different Latin letters to spell the same Chinese characters. In mainland China, Pinyin is the only official system, and names on passports and other official identification must use the Pinyin system.

and find Chinese buyers alone responsible for premiums related to lucky numbers in Chinese superstition and discounts attributable to unlucky numbers.

The results indicate that Chinese buyers pay a 1.7% premium for properties that include an *8* in the address. We provide evidence that this premium does not reflect unobserved quality of the underlying property as Chinese sellers do not command a premium for properties with an *8* in the address. On the other hand, we find mild evidence that Chinese buyers pay a 1.2% discount for addresses that end in a *4*. These results provide the first evidence that Chinese superstitions impact transaction prices in an American real estate market, and indicator of the extent of cultural assimilation.

A falsification test finds no evidence that Korean buyers pay a premium for homes with addresses containing an *8*. Robustness tests adding names from Hong Kong and Taiwan with different Romanization systems also support our findings. In the context of cultural assimilation, we find evidence that Chinese preferences for specific numbers are durable and long-lived, even for minority residents in a city with a multiplicity of cultural preferences and backgrounds. Our results indicate that Chinese superstition is portable and still relevant among Chinese living in America.

2 Literature Review

2.1 Superstition and Real Estate

Previous research examined the role of superstition in the market for apartments in Hong Kong and mainland China. Chau et al. (2001) analyze data from Hong Kong and find apartments on floor 8 sell at a 2.5% premium, while apartments on floor *4* do not sell at a significant discount. Shum et al. (2014) analyze data from Chengdu, a provincial capital city in Western China, and find that apartments located on floors ending with an 8 sell in the secondary market at a premium of 235 RMB per square meter (approximately 7%). No price effects are found in the primary market due to a uniform local pricing policy. In addition to price effects, apartments on floors ending in an *8* sold 6.9 days sooner than apartments on other floors, on average. Using individual data, Shum et al. (2014) identify individuals with phone numbers that contain multiple *8*s as superstitious individuals and find these individuals are more likely to buy an apartment on a floor ending with an *8*. Despite evidence supporting the importance of the number *8*, Shum et al. (2014) find no evidence that the presence of the number *4* is associated with any price discount.

Other researchers found price effects attributable to Chinese superstition in countries outside China. Absent identifiers for Chinese individuals, Bourassa and Peng (1999) and

Fortin et al. (2014) compare property prices in census units with a large concentration of Chinese to property prices in other census units. Bourassa and Peng (1999) examine census units in New Zealand and find positive price effects associated with *8*s in census units with a large percentage of Chinese residents; no such effects are found for similar properties in census units with few Chinese residents. Similar to Bourassa and Peng (1999), Fortin et al. (2014) compare property prices in census units with a large numbers of Chinese to property prices in other census units in the Canadian city of Vancouver, British Columbia. Fortin et al. (2014) find houses with addresses ending in an *8* sell at a 2.5% premium in the census units with many Chinese residents; in the same census units, addresses that end in a *4* sell at a 2.2% discount. No price effects are found in census units with relatively few Chinese residents for either number.

Rehm et al. (2017) analyze residential property transactions in Auckland, New Zealand and identify Chinese buyers using a list of Chinese surnames developed by Quan et al. (2006) and a "sense check" of 100 random transactions from their data by a fluent Mandarin speaker. The "sense check" indicated that 89 out of 100 transactions identified as involving a Chinese buyer based on the surname list were deemed to be Chinese. Rehm et al. (2017) estimate an hedonic model and report a 1.8% premium paid by Chinese buyers for houses with an address containing an *8* over the period 2003-2006 (about 4,000 transactions) and no premium for houses with an address containing an *8* over the period 2011-2015 (about 17,000 transactions). Rehm et al. (2017) find no discount associated with addresses containing a *4*.

Although Bourassa and Peng (1999) and Fortin et al. (2014) provide suggestive evidence of impacts of Chinese superstition on real estate prices outside China, their price effects must be attributed to the property and not the individual buyer or seller. Absent any information on the ethnicity of buyers and sellers, these studies can not identify any individual-level effects attributable to Chinese buyers or sellers. In contrast, Agarwal et al. (2016) and Rehm et al. (2017) identify individual-level effects in real estate markets using explicit Chinese identifiers present in the data. Both find that Chinese buyers pay a premium for homes with addresses ending in *8*; Agarwal et al. (2016) find a discount for apartments with numbers ending in *4*. Similar to Agarwal et al. (2016) and Rehm et al. (2017), we identify individual-level effects, but unlike these papers we impute ethnicity based on a supervised learning approach.

In addition to real estate markets, empirical research has also found Chinese superstition effects in other markets. Woo et al. (2008) and Ng et al. (2010) find evidence using winning bids for license plate auctions in Hong Kong. Yang (2011) document that retailers in China manipulate patterns of numbers appearing on price tags in order to exploit preferences for lucky and unlucky numbers. Moreover, Yang (2011) conclude that Chinese consumers pay

more for retail goods because of this manipulation.

2.2 Name-Ethnicity Matching

In addition to testing for the effect of cultural preferences on real estate prices, this study develops a binomial classifier for placing individuals into specific ethnic groups based on name. The need for a name-ethnicity classification scheme is more practical than ideal, and has historically been based on data available to researchers in the social and biomedical sciences. Treeratpituk and Giles (2012) put this concisely

> unlike names, ethnic information is often unavailable due to practical, political or legal reasons. (page 1142)

Like most empirical real estate research, we use data from the King County Assessor that includes buyer and seller names but does not include ethnic identifiers.

Motivated by genetic commonalities within ethnic groups, name-ethnic matching has been used extensively in biomedical research (Coldman et al., 1988; Burchard et al., 2003; Fiscella and Fremont, 2006). A typical approach identifies strong predictors of ethnicity using a labeled data set that includes both the ethnicity and name for each individual. Coldman et al. (1988) use death certificates that include name and ethnicity, Gill et al. (2005) use surnames and country of origin, and Ambekar et al. (2009) use names of famous natives obtained from Wikipedia. Nowak and Sayago-Gomez (2017) use this approach to identify Arab names in a similar setting.

In this study, we use Olympic Games rosters for both the United States and China from 1948 to 2012 as a representative list of names from each country. Olympic Games team rosters contain both males and females, and the team members must meet specific residence and citizenship requirements in order to appear on the national team for each country. These features makes Olympic Games team rosters an ideal choice for developing representative lists of names by country when compared to other potential labeled data sets such as Wikipedia or the Internet Movie Database, Ambekar et al. (2009) and Rachevsky and Pu (2011).[6][7]

As names are a specific form of textual data, our method relates to other studies that view text as data. We use a tokenization approach where units of text are represented by exchangeable collections of words or *tokens*. Based on the set of tokens, each text can be scored or classified into two or more groups. For example, Gentzkow and Shapiro (2010) score news outlets as Republican or Democrat, Loughran and McDonald (2011) score 10k filings as positive or negative, and Nowak and Smith (2016) score real estate listings as low quality or

[6] https://en.wikipedia.org/wiki/Wikipedia:Database_download
[7] http://www.imdb.com/interfaces

high quality. In order to score the text, researchers can either use a pre-specified dictionary of topic-specific words or build a dictionary based on a corpus of labeled or unlabeled text. We create a dictionary of Chinese and non-Chinese names using sparsity-inducing methods similar to Taddy (2013). Using the names and estimated weights, each name in the assessor data can be scored as either Chinese or not Chinese.

The purpose of the classification procedure is to predict ethnicity for names in the assessor data. Because of this, the performance of the classifier should not be evaluated on in-sample mis-classification for the Olympic Games rosters; rather, performance should be evaluated based on theoretical results for the out-of-sample mis-classification rate of the assessor data. Given the number of unique names in the Olympic Games rosters is comparable to the number of Olympians, overfitting is likely a problem. Because of this, we use an ℓ_1 regularized logistic regression commonly used in the document classification literature, Hastie et al. (2015). Regularizing the coefficients using the ℓ_1 norm yields coefficient estimates that result in lower out-of-sample mis-classification compared to un-regularized estimators and alternative ℓ_p coefficient regularizations (Ng, 2004). Furthermore, unlike the maximum likelihood estimator, the regularized estimator is feasible even when the data are separable (Hastie et al., 2015).

3 Data and Methodology

We estimate an hedonic price model in order to explain observed variation in residential real estate transaction prices in Seattle (King County), Washington attributable to the presence of lucky or unlucky numbers in the address. The hedonic model contains indicator variables for individual buyers and sellers classified as Chinese. We classify based on name using the rosters of the athletes on the Chinese and US Summer Olympic Games over a 60 year period. The data sources and estimation methods used are described in detail below.

3.1 Data

The data sets used in this study come from two sources. The first data set includes the rosters of all Summer Olympic Games teams from the United States and China beginning 1948 and ending 2012. These data form the basis for the supervised learning algorithm used to identify individuals as Chinese, as described below. The Summer Olympic rosters were downloaded from the Sports Reference website.[8] Figure 1 shows the 100 most common names appearing in the US and China national Olympic teams over the 1948-2012 period. In Figure 1, the

[8] http://www.sports-reference.com/olympics/

larger the font size, the more frequently that name appears on the Summer Olympic Games team rosters.

The second data set comes from the King County Assessor's Office.[9] This data set includes information on all real estate transactions in King County beginning January 1, 1990 and ending December 31, 2015. The data set includes information about the property (type of property, type of transaction, address, etc.), the transaction price, the buyer name, and the seller name. We use data on sales of single-family homes. After removing 1% of outlying observations based on a preliminary hedonic regression, the final sample contains 508,916 single family home sales.[10] Summary statistics for commonly reported property attributes are reported in Table 2. The average residential property transacted during the sample period was built in 1978, had a price of $330,555, just under 2,000 square feet of living space, 3.3 bedrooms and about 1.5 bathrooms.

We identify individuals as having a Chinese cultural or ethnic background based on name using a binomial classifier. After training the classifier using names on Olympic Team rosters, we calculate the probability that a given buyer's name will be found on the Chinese Olympic team rosters, $\Pr(ChinaBuyer)$. Using this probability, we create an indicator variable *chinaBuy* which is equal to 1 if $0.8 < \Pr(ChinaBuyer)$ and equal to 0 otherwise. Alternative cutoff values for this indicator variable were considered, but changing the threshold probability across values in the set $\{0.55, 0.60, ..., 0.90, 0.95\}$ did not alter the empirical results in any meaningful way. The probability $\Pr(ChinaSeller)$ and indicator variable *chinaSell* are created in a similar manner using seller names.

Summary statistics for the probabilities, indicator variables, and the presence of *8*s and *4*s in addresses, are shown on Table 2. 4.3% of all buyers are classified as having a name suggesting a Chinese cultural background and 1.9% of all sellers are classified as such. About 33% of the houses in the sample have an *8* in the address, and about 45% have a *4* in the address. About 9% of the homes transacted in the sample have a *4* or *8* as the final digit in the house price.

3.2 Binomial Classifier

For each $n = 1, ..., N$, define an indicator variable $y_n = 1$ if the Olympic athlete is on the Chinese national team and $y_n = 0$ if the Olympic athlete is on the US national team. Using this binary variable, the probability that an Olympic athlete will be from China is calculated

[9] http://www.kingcounty.gov/depts/assessor.aspx

[10] Based on deed records available on the King County Assessor's website, a significant portion of the outlying transactions were found to be associated with non arms-length transactions, inter-family transfers, fire damage, or significant renovation.

using a logit function. Because of the binary nature of the dependent variable, we consider this a binomial classifier.

The explanatory variables for the logit model are created from the full names on Olympic Games team rosters. We assume each full name, F_n, can be represented as an exchangeable collection of names or tokens chosen from a set of P names. The exchangeability assumption implies that we make no distinction between first and last names. Alternatively, each full name F_n can be represented as a $P \times 1$ vector X_n with elements X_{np}. Here, $X_{np} = 1$ if the p^{th} name is in F_n and $X_{np} = 0$ otherwise. For instance, the associated vector X_n for American Olympic swimmer $F_n = \{Michael, Phelps\}$ has a 1 in the element associated with $Michael$, a 1 in the element associated with $Phelps$, and 0 everywhere else. Using these explanatory variables, the probability that $y_n = 1$ is given by

$$\Pr(y_n = 1|X_n, \phi) = \frac{e^{\phi_0 + \sum_p X_{np}\phi_p}}{1 + e^{\phi_0 + \sum_p X_{np}\phi_p}} \tag{1}$$

In Equation (1), when $0 < \phi_p$ ($\phi_p < 0$), the presence of name p increases (decreases) the likelihood that F_n comes from the Chinese Olympic team roster. When $\phi_p = 0$, name p does not help to predict y_n. The parameter ϕ_0 controls the unconditional $\Pr(y_n = 1)$.

For fixed P, ϕ_p can be consistently estimated using maximum likelihood estimation. In the Olympic Roster setting, the assumption of fixed P is difficult to defend as there are 6,502 unique names across $N = 9,836$ Olympic athletes from both the US and China. For sets of explanatory variable with these dimensions, maximum likelihood solutions are at worst degenerate when $N < P$ and at best unreliable when $P \approx N$ (Hastie and Qian, 2014). A practical approach decreases P by filtering out names that occur fewer than C times in the data. In this case, modest filtering rules result in a large P while more aggressive filtering rules could remove names with significant predictive power. We retain the $P = 615$ names that occur $C = 5$ or more times in the data. In unreported results, we find that the results are not sensitive when using $C = 10$ or $C = 20$.

Because P remains large even after filtering out less common names, we utilize a penalized likelihood procedure that mitigates overfitting. In particular, we place an ℓ_1 penalty on the individual ϕ_p parameters and minimize the following penalized likelihood function

$$-\prod_n \Pr(y_n = 1|X_n, \phi)^{y_n} \left[1 - \Pr(y_n = 1|X_n, \phi)\right]^{1-y_n} + \lambda \sum_p |\phi_p| \tag{2}$$

The first term in Equation (2) is the sample likelihood, and the second term represents a penalty on the coefficient vector. The parameter λ is a tuning parameter that controls the

penalty.[11][12] Define the solution to Equation (2) as $\phi^*(\lambda)$. When the context is clear, we omit the dependence on λ and write ϕ^* instead.

The choice of λ determines the size of the penalty on ϕ. When $\lambda = 0$, there is no penalty on ϕ and $\phi^*(0)$ is the maximum likelihood estimator. As λ increases, there is a greater penalty on large ϕ and ϕ^* is shrunk towards the zero vector. Unlike the ℓ_2 penalty, the shape of the ℓ_1 penalty yields a sparse solution where some entries of ϕ^* can be set equal to 0. As mentioned above, when $\phi^*_p = 0$, token p cannot be used to classify y_n. With this interpretation, minimizing Equation (2) performs both variable selection and coefficient estimation. If we were to forgo the logit model and instead estimate a linear probability model with an ℓ_1 penalty on the coefficients, the estimator would become the well-known LASSO estimator, Tibshirani (1996).[13]

By including the penalty term, ϕ^* is less likely to overfit the data in-sample and can be used for meaningful out-of-sample classifications (Ng, 2004). For this application, out-of-sample performance (mis-classification) is fundamental to the results. By creating explanatory variables for buyer and seller names in the assessor data in the same way, we can then calculate $\Pr(ChinaBuyer)$ and $\Pr(ChinaSeller)$ and the associated indicator variables using ϕ^* and Equation1.

In addition to out-of-sample considerations, we also prefer the regularized estimator based on the configuration of the data. $P = 615$ names may be sufficiently small compared to $N = 9,836$ Olympians to justify use of maximum likelihood methods. However, many names are specific to either the Chinese or US rosters. For instance, *michael* is only found on US team rosters. In this case, the data are considered *separable*, and the maximum likelihood estimator does not exist as $\phi^*_{michael}(0) = -\infty$.[14] However, separable data sets can still be employed using the regularized estimator in Equation (2) as $0 < \lambda$ precludes infinite values for ϕ^*.

[11] In our analysis, we experiment with values near the 5-fold cross-validated λ. The results are robust to λ near the cross-validated choice of λ

[12] We use the `glmnet` package in `R` to solve Equation 2. The solution is found by using a quadratic approximation to the true penalized likelihood.

[13] LASSO is an acronym for *least absolute shrinkage and selection operator*.

[14] Using Eq 1, the individual likelihood of a Chinese Olympian, $y_n = 1$, is not affected by $\phi_{michael}$ as $x_{n,michael} = 0$ for all Chinese Olympians. The maximum likelihood estimator maximizes the sample likelihood $L(\phi) = \prod_n \Pr(y_n = 1|X_n, \phi)^{y_n} [1 - \Pr(y_n = 1|X_n, \phi)]^{1-y_n}$. Suppose ϕ^1 maximizes $L(\phi)$. Now, consider $\phi^2 = \phi^1 - e_{michael}$ where $e_{michael}$ is the basis vector with a 1 in the *michael* slot and 0 elsewhere. Using Eq 1, $\Pr(y_n = 1|X_n, \phi^2) = \Pr(y_n = 1|X_n, \phi^1)$ for all Chinese Olympians and US Olympians not named *michael*, $\Pr(y_n = 1|X_n, \phi^1) < \Pr(y_n = 1|X_n, \phi^2)$ for all US Olympians names *michael*. Therefore, $L(\phi^1) < L(\phi^2)$, and ϕ^1 cannot be the maximum likelihood estimator.

3.3 Multinomial Classifier

As discussed above, different Chinese and Korean names can be identical when Romanized. In this situation, the binomial classifier might erroneously classify Korean buyers and sellers as Chinese. Using a multinomial classifier instead of a binomial classifier in this setting provides two benefits. First and foremost, adding a third type can decrease classification error relative to the binomial classifier. Specifically, Romanized names that are common to both China and Korea will not default to being classified as Chinese as in the binomial classifier. Second, extending the classification scheme by allowing for a Korean type also provides for an interesting falsification test. Unlike Chinese superstition, there does not exist any evidence that the number *8* is lucky or unlucky in Korean superstition.

The binomial classifier described in the previous section can be generalized to a multinomial classifier using a multinomial likelihood approach. The multinomial classification model contains $k = 1, ..., K$ types. Each individual $n = 1, ..., N$ is associated with a type $y_n \in \{1, ..., K\}$. Given the vector X_n, the probability of being type k is given by

$$\Pr(y_n = k|X_n, \phi) = \frac{e^{\phi_{0k}+X_n'\phi_k}}{\sum_k e^{\phi_{0k}+X_n'\phi_k}} \tag{3}$$

In Equation (3), $\phi_k = (\phi_{1k}, ..., \phi_{Pk})'$ is a $P \times 1$ vector of parameters for type k. When $0 < \phi_{pk}$ ($\phi_{pk} < 0$), the presence of name p increases (decreases) the likelihood that F_n is type k. In the interest of out-of-sample performance, ϕ_k can be estimated by maximizing a penalized likelihood similar to Equation 2. Using ϕ_k^*, we calculate probabilities $\Pr(ChinaBuyer)$ and $\Pr(KoreanBuyer)$ and create the indicators *chinaBuy* and *koreaBuy* using the same 0.8 cutoff used in the binomial classifier. Indicators for sellers are created similarly.

3.4 Hedonic Price Model

In order to isolate the response of Chinese buyers and sellers to the presence of certain numbers, we use the property address recorded in the King County Assessors Office real estate transactions database. In these data, the property address includes both the building number and street number. For example, *248 Main Street* has a single *8* in the address while *248 8th Street* has two *8*s in the address. In the transaction data, we convert all character representations of numbers to numerics. For instance, *248 Eighth Street* is converted to *248 8th Street.*

Indicator variables for the presence of *8*s and *4*s are created using the property address. The variable *any*8 = 1 if there is any *8* in the property address and *any*8 = 0 otherwise. The variable *total*8 is equal to the total number of *8*s in the property address. In order to

determine if the building number and street have different effects, we set $buildingAny8 = 1$ if the building number contains an *8* and $buildingAny8 = 0$ otherwise. Following Fortin et al. (2014), we also create $buildingLast8 = 1$ if the last digit of the house number is equal to 8 and set $buildingLast8 = 0$ otherwise. Indicators for the number *4* are created in a similar manner. As an example, a single family home at 248 8th Street would have *any8=1*, *buildingAny8=1*, *buildingLast8=1*, *total8=2*, *any4=1*, *buildingAny4=1*, *buildingLast4=0*, and *total4=1*.

We estimate a hedonic model in order to determine if individuals with a Chinese cultural background are willing to pay more or less for a single family home based on the numbers found in the address. We estimate the following hedonic price model for house i in zip code z sold at time t

$$p_{izt} = x_{izt}\beta + \psi z_{izt} + \mu_{zt} + u_{izt}. \tag{4}$$

In Equation (4), p_{izt} is the log of the sale price, x_{ict} includes the log square footage, bedrooms, bathrooms, and age of the property, z_{izt} includes indicator variables for Chinese (*chinaBuy*, *chinaSell*), numbers appearing in the street address (*any*8, *buildingAny*8, etc.), and the relevant interaction terms (*any*8 × *chinaBuy*, *any*8 × *chinaSell*, etc.), μ_{zt} is a Zip Code - Year fixed-effect that captures time-varying unobservable neighborhood heterogeneity, and u_{izt} is an unobservable error term capturing other factors that affect residential property transaction prices. We two way cluster-correct the estimated standard errors in Equation (4) at the Zip Code and year level.

In Equation (4), the coefficients for *chinaBuy* and *chinaSell* indicate if Chinese buyers or sellers, respectively, pay more or less for residential properties regardless of the numbers in the address. When the coefficient on *chinaBuy* is positive, Chinse buyers pay a premium when purchasing a residential property. Of course, as the hedonic model will never fully capture the true quality of a property, and a positive coefficient on *chinaBuy* can also indicate that individuals with a Chinese background purchase properties with higher unobserved quality.

Our primary variables of interest are the interaction terms like *any*8 × *chinaBuy*. The coefficient on *any*8×*chinaBuy* indicates any premium or discount Chinese pay when purchasing properties with any *8*s in the property address. This premium or discount is attributable solely to the numbers in the property address and, by the inclusion of *chinaBuy* as a standalone coefficient, is in addition to any market wide premium paid by Chinese buyers. If Chinese buyers factor in Chinese superstition when purchasing a property, we expect the coefficient on *any*8 × *chinaBuy* and other interaction terms that include *chinaBuy* and *8*s to be positive.

If properties that include *8*s purchased by Chinese are of higher quality, a positive coeffi-

cient on $any8 \times chinaBuy$ could indicate unobserved quality and not the influence of Chinese superstitions. Estimates of the coefficient on the interaction term $any8 \times chinaSell$ can help address this problem. If the estimated coefficient on $any8 \times chinaSell$ is not different from 0, this is strong evidence that properties sold by Chinese with an *8* in the address are not of higher or lower quality than other properties. Altogether, both a statistically positive coefficient estimate on $any8 \times chinaBuy$ and a coefficient estimate indistinguishable from zero on $any8 \times chinaSell$ indicates Chinese buyers paying a premium for properties based solely on the presence of an *8* in the address irrespective of the unobserved quality of the property.

In contrast to Bourassa and Peng (1999) and Fortin et al. (2014) who interact an indicator variable for census units that include a large portion of Chinese, say $chineseTract$, with indicators for *8* in the address, we identify Chinese buyers and sellers.[15] This subtle difference is important if non-Chinese recognize the effects of Chinese superstitions and purchase properties for speculative purposes. That is, a positive coefficient on $any8 \times chineseTract$ can indicate a positive price effect for the number *8* for both Chinese and non-Chinese alike. In contrast, a positive coefficient on $any8 \times chinaBuy$ identifies a positive price effect specific to Chinese buyers.

4 Results

4.1 Ethnic-Name Matching

The ethnic name matching procedure is a key element of the empirical analysis. Figure 2 shows the solution path, $\phi^*(\lambda)$, for various values of λ in Equation (2). Figure 2 shows both the size of the coefficients and the number of non-zero coefficients increasing as λ decreases. Furthermore, Figure 2 indicates that each coefficient becomes non-zero at different values of λ. For large values of λ, only the names that are the strongest predictors have non-zero ϕ^*. Therefore, the choice of λ directly determines both the dictionary of Chinese and US names (variable selection) and the predictive power of the names (coefficient estimation).

The choice of λ is determined using a cross-validation procedure. λ_{cv} is the 5-fold cross-validated choice of λ, and λ_{1se} is the largest λ such that the median cross-validated log-likelihood is within one standard error of the log-likelihood evaluated at λ_{cv}. Although some practitioners favor using $\phi^*(\lambda_{1se})$ as it is more conservative in terms of both variable selection and coefficient estimation (Hastie et al., 2015), parameter estimates from Equation (4) are nearly identical when using either λ_{cv} or λ_{1se}. Therefore, we report only the results using

[15] Bourassa and Peng (1999) identify tracts based on immigration and Fortin et al. (2014) identify tracts based on census data.

$\phi^*(\lambda_{cv})$, hereafter, ϕ^*_{cv}.

Estimates of ϕ^*_{cv}, the name matching parameter from Equation (1), along with the largest estimated values are displayed in Table 1. Names that most strongly predict being on the United States Olympic team roster are *kevin, amy, michael*. Names that most strongly predict being on the Chinese Olympic team roster are *li, yin, xu*. Using Equation 1, the implied $\Pr(y_n = 1)$ for *kevin, amy, michael* is equal to 0 when rounding to 6 digits; the implied $\Pr(y_n = 1)$ for *li, yin, xu* is equal to 1. Thus, the presence of any of these names alone in any buyer or seller name is a strong indicator of ethnicity.

Not surprisingly, the strongest predictors are names that are among the most frequent names in Figure 1. However, there is not a monotonic relationship between frequency and predictive power. For instance, *dan, lou, lee, long* are found on both Chinese and United States Olympic team rosters; *dan* occurs 17 (39) times in the United States (Chinese) Olympic rosters and is not a strong indicator of ethnicity. As mentioned above, the ℓ_1 penalty in Equation 2 is such that ϕ^*_{cv} for weak predictors are set exactly to 0. The associated ϕ^*_{cv} for the 15 names that occur in both the Chinese and US rosters are equal to 0 indicating that these 15 names cannot be used to predict ethnicity in the assessor data.

Using ϕ^*_{cv}, the probability that a buyer or seller is Chinese can be calculated using Equation 1. In the data, 3.5% of the transactions have a buyer name with $0.95 < \Pr(ChineseBuyer)$. A manual inspection of the names by several Chinese nationals confirms this high predicted probability. As indicated in Table 2, 4.3% of transactions are classified as involving Chinese buyers, and 1.9% of transactions involve Chinese sellers. Figure 3 shows the fraction of transactions that included either a Chinese buyer or seller over the sample period. The fraction of Chinese buyers increased at a steady rate beginning in 1990 through 2008. After 2008, the fraction of Chinese buyers increased more rapidly, peaking at more than 8% of all buyers in 2013. In contrast, the percentage of Chinese sellers exhibits more steady growth rate throughout the sample period.

The percentage of Chinese buyers and sellers varies across locations in Seattle. Figure 4 shows the fraction of Chinese buyers in King County by census tract and Figure 5 shows the locations of the individual properties associated with Chinese buyers. Although transactions involving Chinese buyers are distributed throughout King County, significant clusters of transactions involving Chinese buyers can be seen on Figure 5. The fraction of Chinese buyers appears to be highly concentrated in two locations where more than 20% of buyers are identified as Chinese by the binomial classifier. Similar high home ownership rates and clustering patterns among Chinese is also reported by Painter et al. (2004) in their study using data from the Los Angeles Consolidated Metropolitan Statistical Area.

One cluster of Chinese buyers is in the Beacon Hill area of Seattle just east of I-5

and the Seattle-Tacoma International Airport. It is interesting to note that the Chinatown International-District is located 3 miles north of the Beacon Hill area.[16] The other location is the Newcastle / Cougar Hills area south of I-90 and east of I-405.

The names with the largest coefficients for each country in the multinomial classifier are presented in Table 4. Not surprisingly, the strongest predictors for Chinese names in the multinomial model are comparable to the strongest names in the binomial model given in Table 1. The total number of buyers and sellers for each type are presented in Table 5. The total number of Chinese buyers and sellers in Table 5 is fewer than the total numbers in Table 3. A comparison of the counts in the two tables indicates that the binomial classifier is classifying buyers and sellers in the assessor data as Chinese who are classified as Korean when using the multinomial classifier.

4.2 Hedonic Results for the Binomial Classifier

We use four alternative specifications for the hedonic model in Equation 4 that contain different indicator variables for the presence of lucky and unlucky numbers in addresses in different forms. These alternative models help establish the robustness of the results. Model 1 contains *any8* or *any4* and interactions with *chinaBuy* and *chinaSell*. Model 2 uses *total8* or *total4*. Model 3 uses *buildingAny8* or *buildingAny4*. The final model specification, Model 4, contains an indicator variable (*buildingLast8* or *buildingLast4*) for the presence of an *8* or *4* as the last digit of the house number. This specification in Model 4 matches the one used by Fortin et al. (2014). We interact the indicator variables for lucky and unlucky numbers in addresses with indicator variables for Chinese buyers and sellers, which allows for the effect of Chinese superstition to vary depending on which party in the transaction has these preferences.

The results for the hedonic regression model defined by in Equation (4) are presented in Table 6.[17] All models contain indicator variables for transactions with Chinese buyers and sellers. The estimated parameters on these stand-alone indicator variables are all negative and statistically different from zero; Chinese buyers and sellers in King County tend to purchase single family homes at a discount relative to other buyers. This price effect could reflect a preference of Chinese for time-invariant low-quality properties. Alternatively, this price effect could also reflect the time-varying condition of the property or features specific

[16] http://www.visitseattle.org/visitor-information/

[17] One potential concern is that the two-step estimation process leads to bias in the second step imputed regressors and standard errors, which reflect first step sampling error (Murphy and Topel, 1985). Therefore, we bootstrap both the Olympic roster model and Seattle housing transactions model and re-estimate the coefficients and standard errors. Results are very comparable with those reported. For more details, please see Appendix A.

to the transacting parties. For example, Chinese could purchase high-quality properties that are currently in poor condition.[18] The discount could also be an indication that Chinese have bargaining power relative to the average seller, Harding et al. (2003). However, we find the bargaining power explanation unlikely as the coefficient on *chinaSell* is also negative and smaller than the coefficient on *chinaBuy*.

The results from Model 1 suggest no significant discount for a property address with an *8* in the address across King County. The parameters of interest are the estimates on the interaction of *any*8 and with the indicators for Chinese buyers and sellers. The interation parameter estimates on Table 6 indicate that Chinese buyers pay a 1.7% premium for property addresses that include an *8*. Chinese sellers receive a 1.4% premium when selling a property with an *8* in the address, no matter what the ethnicity of the buyer. The premium when selling is puzzling but could represent either model mis-specification or mis-classification of Chinese buyers and sellers when using the binomial classifier. The premium when selling could also indicate a larger reservation price for Chinese who currently derive utility from properties than include an *8* in the address.

Columns 2 - 4 on Table 6 investigate alternative specifications for the presence of *8*s in addresses. Model 2 includes a variable reflecting the total number of *8*s in the property address. Again, the presence of an *8* in an address does not carry any premium in the overall sample. However, Chinese home buyers are willing to pay a 1.4% premium for each additional *8* in a home's address. Chinese sellers receive a 0.9% premium for each additional *8* in an address; however, this result is only significant at the 5% level. Model 3 includes an indicator variable for the presence of an *8* anywhere in the building address. Results from Model 3 indicate that buyer and seller premiums for *8*s appearing in the house number are similar to the premium for an *8* anywhere in the address.

Model 4, shown on in column 4 on Table 6, contains an indicator variable for houses where the final digit of the house number is *8*. The average single family home transaction in King County involving a property with an *8* as the final digit of the house number does not carry any premium when compared to other properties. However, results in Table 6 echo the same puzzling result from Model 1 where Chinese sellers command a 2.1% premium.

Table 7 presents a similar analysis using *4*s in addresses. Again, Chinese buyers purchase single family homes for a price 2.0% to 2.4% below average and sell single family homes at a price about 5.3% below average in King County. Similar to Fortin et al. (2014), Models 1-3 find no evidence that Chinese buyers react to the presence of *4* in the address. However,

[18] Here, we use *quality* to define the time-invariant or slowly-varying state of the property (location to amenities, neighborhood quality, school district, etc.) and *condition* to indicate the time-varying state of the property (fire damage, flood damage, renovated kitchen, new roof, etc.).

Model 4 indicates Chinese buyers pay a 1.2% discount when the last digit of the house number is a *4*.

The results from the binomial classifier suggest that single family home transaction prices in Seattle, Washington reflect cultural preferences for lucky and unlucky numbers. King County has a diverse population that included about 15% of the population identifying themselves as Asian in the 2010 Census. This is a substantially more diverse ethnic mix than the setting examined by Agarwal et al. (2016) and Shum et al. (2014), who analyze the premium (discount) associated with the presence of *8* (*4*) in majority Chinese settings. Chinese in Seattle interact more frequently with people from a western background than residents of China or Singapore, and are also continually bombarded by media with a western orientation. Some Seattle residents identified as Chinese could be second, third forth or more generation Chinese-Americans. These results suggest that Chinese cultural preferences for specific numbers persist over time, and in the presence of significant interaction with, and exposure to non-Chinese cultural preferences.

The estimated premia associated with the presence of *8*s in addresses, and the estimated discount associated with the presence of *4*s in addresses, in this paper are smaller than those reported in Fortin et al. (2014), and substantially smaller than those reported in Shum et al. (2014). The data used by Shum et al. (2014) come from a city in China, where cultural preferences for numbers should be substantially stronger than in Seattle. Fortin et al. (2014) have no information about the ethnicity of buyers and sellers; instead, they exploit information about the demographic characteristics of the Census Tracts where the houses are located in Vancouver.

4.3 Hedonic Results for Multinomial Classifier

We next present results when using the multinomial classifier to classify buyer and seller ethnicity. Results for Chinese buyers are presented in Table 8, and results for Korean buyers are presented in Table 9. The results for the Chinese buyers and sellers are comparable to the results in Table 6; the results from the previous section are robust to use of the multinomial classifier. We find further evidence that Chinese buyers are willing to pay a slight premium for properties that include an *8* in the address. Interestingly, the puzzling, statistically significant result for the $buildingLast8 \times chinaSell$ variable in Table 6 is no longer significant in Table 8. The disappearance of this significance indicates that the multinomial classifier is preferred to the binomial classifier in terms of its ability to assign ethnicity to buyers and sellers in the sample.

More importantly, as expected, we find that Korean buyers do not pay a premium for

properties that include an *8* in the address. As a whole, the null result for Koreans in Table 9 and the significance for Chinese buyers in Table 8 provide two important implications. First, there is evidence that Chinese buyers factor in Chinese superstition and are willing to pay more for properties that include *8*s in the address. This result is robust to any mis-classification attributable to the Romanization of Chinese characters. Second, the multinomial classifier generates a simple counterfactual: the absence of any significant price effects for Korean buyers and sellers gives further credence to the significant results found among Chinese buyers and sellers in the sample.

4.4 Robustness Tests

Although the Pinyin system has been widely used to romanize Chinese names, concerns remain that some house buyers and sellers are from Hong Kong and Taiwan, where Pinyin is not the only romanization method used. The current official system in Hong Kong is the "Hong Kong Government Cantonese Romanisation", which uses Cantonese dialects when creating the romanized forms for Chinese characters.[19] Taiwan has used Wades-Giles as the de facto romanization standard for decades. Several other official romanization methods were also used in Taiwan, but all of them are obscure. Since 2008, the Taiwan government has promoted the use of the Pinyin system, but use of the Pinyin system is suggested, not required.[20]

In addition, some Chinese might have immigrated into the U.S. decades, or even more than a century ago, and those from families living in the U.S. for several generations may keep using romainzed names from non-Pinyin systems, most likely the Wade–Giles system. Because of these issues, we need to address threats to classification of names into the "Chinese" group caused by omitting romanized names used in Hong Kong (mainly Cantonese romanizations) and Taiwan (mainly Wade-Giles romanizations). To address these problems, we implement two different robustness tests.

First, we include the names from the rosters of all Summer Olympic Games teams from China, Hong Kong and Taiwan (Chinese Taipei) beginning 1948 and ending 2012 in the "China" category of the classifier. We then employ the multinomial classifier to identify the most "predictable" names for China (including Mainland China, Hong Kong and Taiwan), Korea, and United States. The names with the largest coefficients for each "country" in this multinomial classifier are shown in Table 10.

The strongest predictors for Chinese names are comparable to those from the binomial model shown in Table 1 and those from the multinomial approach shown in Table 4. The

[19] https://en.wikipedia.org/wiki/Hong_Kong_Government_Cantonese_Romanisation
[20] https://en.wikipedia.org/wiki/Wade-Giles

total number of buyers and sellers for each type are shown in Table 11. The total number of Chinese buyers and sellers in Table 11 are larger than those in Table 5, because Table 11 includes other Chinese whose names are not based on the Pinyin system, who were not classified as Chinese by the former classification procedure. A comparison of the counts of the Korean buyers and sellers in the two tables indicates that some Chinese buyers and sellers might have been classified as Koreans when non-Pinyin Chinese names were omitted from the training data.

Regression model results for this alternative definition of Chinese buyers are shown in Table 12, and results for Korean buyers in Table 13. The results for the Chinese buyers and sellers are comparable to those in Tables 6 and 8. Including names from Hong Kong and Taiwan generates similar evidence of Chinese buyers and sellers paying a premium for the number *8* in an address.

However, here we find evidence of superstition effecting Chinese sellers. The $total8 \times chinaSell$ and $buildingAny8 \times chinaSell$ in Table 8 do not have statistically significant parameter estimates, while both are significant in Table 12. The significant $buildingLast8 \times chinaSell$ parameter in Table 8 becomes insignificant in Table 12. Again, like the results in Table 8, all $any8 \times chinaBuy$, $total8 \times chinaBuy$, and $buildingAny8 \times chinaBuy$ have significant positive effects, at the 0.1% significance level. More importantly, Table 13 shows that Korean buyers do not pay any premium for a number *8* in the address, regardless of the form. Therefore, Tables 12 and 13 jointly support our findings of a premium to the number *8* in the address by Chinese buyers only.

The significant results found by adding Olympic players' names from Hong Kong and Taiwan to the "Chinese" category could be caused by strong effects from only buyers from Mainland China. Or, buyers from Hong Kong and Taiwan, and those whose family immigrated into Unites States several generations ago, might have different superstitious beliefs. Thus, we implement another multinomial classifier analysis, using three types of names: "Chinese", including Olympic roster names from only Mainland China; "hktw", including those from Hong Kong (HK) and Taiwan (TW); and United States. The names with the largest coefficients for each "type" in this multinomial classifier are presented in Table 14. The strongest predictors for Chinese names are comparable to those in the binomial model given in Table 1 and those in the former multinomial approaches given in Tables 4 and 10.

Table 14 also shows the most "predictable" names for Hong Kong and Taiwan.[21] The

[21] The majority of the names are Chinese ones from the Hong Kong and Taiwan Olympic rosters. However, there are some exceptions, because some non-Chinese athletes represented Hong Kong in the Olympic Games. For example, the "rull" in Table 14 is not a Chinese name. Peter Rull Sr. , originally from Estonia, and his son Peter Rull Jr. , represented Hong Kong in several Summer Olympic Games. Robustness checks that manually exclude several non-Chinese names like "rull" from the Hong Kong rosters generate comparable

total number of buyers and sellers for each type are presented in Table 15. As expected, the sum of the number of buyers from Mainland China, Hong Kong, and Taiwan in Table 15 is comparable to that form Chinese buyers in Table 11, as is the sum of the Chinese sellers.

Regression results for Mainland Chinese buyers are presented in Table 16, and results for buyers from Hong Kong and Taiwan are presented in Table 17. The results for buyers both from Mainland China and from Hong Kong and Taiwan are comparable to those generated by earlier binomial and multinomial approaches, and the results for sellers from these Chinese regions are not significant but still comparable to those found using earlier approaches in general. The robustness test results support our findings of the existence of superstition-related premiums by Chinese in the housing market in Seattle.

5 Conclusions

A growing body of evidence suggests that superstition is manifest in economic outcomes. We use a novel approach to identify the ethnicity of home buyers and sellers in King County, Washington over a fifteen year period. The results reveal that the presence of the number *8* in an address has a expected premium for Chinese buyers of 1.7%. A similar premium also exists in other numeric formats, including the total number of *8*s in the address and an *8* as the final digit of the house number. However, the presence of the number *4* in an address does not generate a substantial discount.

The results in this paper extend economists' understanding of the extent of cultural assimilation. Seattle is ethnically and culturally diverse, and many of the ethnic Chinese buying and selling houses during this period could have lived in the U. S. for generations. The presence of a statistically significant relationship between a proxy for the presence of cultural preferences for specific numbers and single family home prices indicates Chinese superstition is relatively durable and still present in Chinese living in America.

In addition, the supervised learning approach to identifying ethnicity based only on names used here can be applied in a number of other settings where quantitative data on language is used. Researchers often use data where names are available but information on ethnicity is not. For example, government regulators, elected officials, political candidates, CEOs, Corporate Board members, judges, and athletes on professional sports teams are often identified by name. However, information about their ethnic background is often limited but is still of significant interest to researchers. The supervised learning approach used here can be applied

results for Mainland Chinese, Hong Kong and Taiwan buyers, and the effects estimated for buyers from Hong Kong and Taiwan are even stronger and more significant than those based on estimation without manual exclusion of non-Chinese names. Results are available upon request.

in all of these settings in order to assess the likely ethnic background of individuals.

References

Agarwal, S., He, J., Liu, H., Png, I. P., Sing, T. F., and Wong, W.-K. (2016). Superstition, conspicuous spending, and housing markets: Evidence from Singapore. *IZA Discussion Paper*, No. 9899.

Ambekar, A., Ward, C., Mohammed, J., Male, S., and Skiena, S. (2009). Name-ethnicity classification from open sources. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 49–58. ACM.

Bourassa, S. C. and Peng, V. S. (1999). Hedonic prices and house numbers: The influence of feng shui. *International Real Estate Review*, 2(1):79–93.

Burchard, E. G., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., Mountain, J. L., Pérez-Stable, E. J., Sheppard, D., and Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, 348(12):1170–1175.

Chau, K. W., Ma, V. S. M., and Ho, D. C. W. (2001). The pricing of "luckiness" in the apartment market. *Journal of Real Estate Literature*, 9(1):29–40.

Coldman, A. J., Braun, T., and Gallagher, R. P. (1988). The classification of ethnic status using name information. *Journal of Epidemiology and Community Health*, 42(4):390–395.

Fernández-Huertas Moraga, J., Ferrer-i Carbonell, A., and Saiz, A. (2017). Immigrant locations and native residential preferences: Emerging ghettos or new communities? *IZA Discussion Paper*, No. 11143.

Fiscella, K. and Fremont, A. M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41(4 pt 1):1482–1500.

Fortin, N. M., Hill, A. J., and Huang, J. (2014). Superstition in the housing market. *Economic Inquiry*, 52(3):974–993.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from US daily newspapers. *Econometrica*, 78(1):35–71.

Gill, P. S., Bhopal, R., Wild, S., and Kai, J. (2005). Limitations and potential of country of birth as proxy for ethnic group. *BMJ: British Medical Journal*, 330(7484):196.

Harding, J. P., Rosenthal, S. S., and Sirmans, C. (2003). Estimating bargaining power in the market for existing homes. *Review of Economics and Statistics*, 85(1):178–188.

Hastie, T. and Qian, J. (2014). Glmnet vignette.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Murphy, K. M. and Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 3(4):370–379.

Ng, A. Y. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM.

Ng, T., Chong, T., and Du, X. (2010). The value of superstitions. *Journal of Economic Psychology*, 31(3):293–309.

Nowak, A. and Sayago-Gomez, J. (2017). Homeowner preferences after September 11th, a microdata approach. *Regional Science and Urban Economics*, (In press).

Nowak, A. and Smith, P. (2016). Textual analysis in real estate. *Journal of Applied Econometrics*, In press.

Painter, G., Yang, L., and Yu, Z. (2004). Homeownership determinants for Chinese Americans: Assimilation, ethnic concentration and nativity. *Real Estate Economics*, 32(3):509–539.

Quan, H., Wang, F., Schopflocher, D., Norris, C., Galbraith, P. D., Faris, P., Graham, M. M., Knudtson, M. L., and Ghali, W. A. (2006). Development and validation of a surname list to define chinese ethnicity. *Medical Care*, 44(4):328–333.

Rachevsky, L. and Pu, K. Q. (2011). Selection of features for surname classification. In *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*, number 4, pages 15–20. IEEE.

Rehm, M., Rehm, M., Chen, S., Chen, S., Filippova, O., and Filippova, O. (2017). House prices and superstition among ethnic Chinese and non-Chinese homebuyers in Auckland, New Zealand. *International Journal of Housing Markets and Analysis*, (In press).

Shum, M., Sun, W., and Ye, G. (2014). Superstition and "lucky" apartments: Evidence from transaction-level data. *Journal of Comparative Economics*, 42(1):109–117.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Treeratpituk, P. and Giles, C. L. (2012). Name-ethnicity classification and ethnicity-sensitive name matching. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1141–1147.

Woo, C.-K., Horowitz, I., Luk, S., and Lai, A. (2008). Willingness to pay and nuanced cultural cues: Evidence from Hong Kong's license-plate auction market. *Journal of Economic Psychology*, 29(1):35–53.

Yang, Z. (2011). "Lucky" numbers, unlucky consumers. *The Journal of Socio-Economics*, 40(5):692–699.

Tables and Figures

Table 1: Olympic Athlete Names and Logit Coefficients

PANEL A: 10 Strongest Predictors for United States Olympians			
Name	Count	Relative Frequency	$\widehat{\phi^*}$
kevin	40	0.004	-6.593
amy	29	0.003	-5.824
michael	67	0.007	-5.514
mike	112	0.011	-5.460
bob	91	0.009	-5.378
jim	89	0.009	-5.327
bill	95	0.010	-5.326
tom	81	0.008	-5.285
steve	72	0.007	-5.184
mark	57	0.006	-5.145
PANEL B: 10 Strongest Predictors for Chinese Olympians			
Name	Count	Relative Frequency	$\widehat{\phi^*}$
li	274	0.028	5.775
yin	10	0.001	5.764
xu	60	0.006	5.712
liu	138	0.014	5.701
sun	61	0.006	5.679
lin	42	0.004	5.468
song	26	0.003	5.239
guo	37	0.004	5.181
yu	52	0.005	5.146
zhu	42	0.004	5.111

Table 1 shows the 10 strongest predictors for Summer Olympic national team members ($\widehat{\phi^*}$s) for the United States and China based on the penalized logit estimator defined by Equation 2. *Count* is the total number of times the name appears on both rosters; *Relative Frequency* is the percentage of times the name appears on both rosters. The strength of the predictor is based on the absolute value of $\widehat{\phi^*}$. Coefficients with more negative (positive) values are strong indicators of a name coming from the United States (Chinese) Summer Olympic team.

Table 2: Summary Statistics

Statistic	Min	Mean	Median	Max	St. Dev.
Sale Price ($1,000s)	45.000	330.555	275.000	1,700.000	208.834
Square Feet of Living Space	480	1,986.760	1,880	4,850	775.857
Year Built	1900	1967.660	1972	2014	27.600
Bedrooms	1	3.328	3	6	0.841
Bathrooms	1	1.498	1	3	0.590
Sale Year	1990	2002.143	2002	2015	6.621
pr(Chinese Seller)	0.000	0.041	0.002	1.000	0.125
pr(Chinese Buyer)	0.000	0.061	0.001	1.000	0.191
chinaSell	0	0.019	0	1	0.136
chinaBuy	0	0.043	0	1	0.203
Any 8 in Address	0	0.332	0	1	0.471
Last Digit 8 in Address	0	0.088	0	1	0.283
Any 4 in Address	0	0.453	0	1	0.498
Last Digit 4 in Address	0	0.096	0	1	0.295

Real estate transaction data comes from the King County Assessor's Office.

Table 3: Number of Identifying Transactions, Binomial Classifier

Variable	Count
Chinese Seller	9,570
Chinese Buyer	21,853
Any 8 in Address (*any*8)	169,182
Last digit 8 in address (*buildingLast*8)	44,748
Any 4 in Address *any*4)	230,520
Last digit 4 in address (*buildingLast*4)	48,966

The Chinese ethnicity indicator variables *chinaBuy* and *chinaSell* are created using the binomial classifier. *any*8 is an indicator for the presence of any *8* in the address. *buildingLast*8 is an indicator if the house number ends in an *8*. *any*4 is an indicator for the presence of any *4* in the address. *buildingLast*4 is an indicator if the house number ends in a *4*

Table 4: Olympic Athlete Names and 10 Largest Multinomial Coefficients

China	$\widehat{\phi^*}$	Korea	$\widehat{\phi^*}$	United States	$\widehat{\phi^*}$
li	8.838	yeong	8.778	kevin	6.872
liu	8.782	cheol	8.773	white	5.873
xu	8.404	choi	8.702	michael	4.329
zhu	8.313	ja	8.523	amy	3.777
zhou	8.273	sin	8.487	david	3.215
xie	8.190	hye	8.286	mike	3.091
he	8.179	won	8.273	ann	3.070
zhao	8.159	seung	8.248	bob	3.011
guo	8.140	seong	8.022	bill	3.010
shen	7.979	yeo	7.604	mary	2.992

Table 4 shows the 10 largest estimated regression coefficients associated with Chinese, Korean, and American names from the multinomial classifier.

Table 5: Multinomial Classifier Transaction Counts

Ethnicity Indicator	Number of Transactions
chinaSell	7,464
chinaBuy	19,287
koreaSell	2,784
koreaBuy	4,495

The ethnicity indicator variables *chinaBuy* and *chinaSell*, *koreaBuy*, and *koreaSell* are created using the multinomial classifier.

Table 6: Buyer and Seller Ethnicity and *8*s in Addresses, Binomial Classifier

	Model 1	Model 2	Model 3	Model 4
chinaSell	-0.055^{***}	-0.054^{***}	-0.053^{***}	-0.052^{***}
	(0.006)	(0.006)	(0.006)	(0.006)
chinaBuy	-0.030^{***}	-0.030^{***}	-0.028^{***}	-0.025^{***}
	(0.007)	(0.007)	(0.006)	(0.006)
any8	0.001			
	(0.003)			
any8 × chinaSell	0.014^{**}			
	(0.005)			
any8 × chinaBuy	0.017^{***}			
	(0.004)			
total8		0.000		
		(0.003)		
total8 × chinaSell		0.009^{*}		
		(0.004)		
total8 × chinaBuy		0.014^{***}		
		(0.003)		
buildingAny8			0.001	
			(0.003)	
buildingAny8 × chinaSell			0.012	
			(0.006)	
buildingAny8 × chinaBuy			0.015^{***}	
			(0.004)	
buildingLast8				0.002
				(0.001)
buildingLast8 × chinaSell				0.018^{**}
				(0.006)
buildingLast8 × chinaBuy				0.003
				(0.004)
Num. obs.	508916	508916	508916	508916
R^2	0.871	0.871	0.871	0.871
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Standard errors cluster corrected at Zip Code-year level. *chinaSell* is an indicator for a Chinese seller, and *chinaBuy* is an indicator for a Chinese buyer. Individuals are classified as either Chinese or non-Chinese using the logit classifier in Equation 1. *any*8 is an indicator for the presence of any *8* in the address. *total*8 is the total number of *8*s in the address. *building*8 is an indicator for the presence of an *8* in the house number. *buildingLast*8 is an indicator for house numbers ending in an *8*.

Table 7: Buyer and Seller Ethnicity and *4*s in Addresses, Binomial Classifier

	Model 1	Model 2	Model 3	Model 4
chinaSell	-0.052^{***}	-0.053^{***}	-0.053^{***}	-0.051^{***}
	(0.007)	(0.007)	(0.007)	(0.006)
chinaBuy	-0.028^{***}	-0.028^{***}	-0.024^{***}	-0.024^{***}
	(0.006)	(0.006)	(0.006)	(0.007)
any4	0.004			
	(0.003)			
any4 × chinaSell	0.004			
	(0.005)			
any4 × chinaBuy	0.007			
	(0.005)			
total4		0.003		
		(0.003)		
total4 × chinaSell		0.004		
		(0.004)		
total4 × chinaBuy		0.005		
		(0.003)		
buildingAny4			0.001	
			(0.002)	
buildingAny4 × chinaSell			0.006	
			(0.004)	
buildingAny4 × chinaBuy			−0.001	
			(0.004)	
buildingLast4				−0.003
				(0.001)
buildingLast4 × chinaSell				0.006^{*}
				(0.003)
buildingLast4 × chinaBuy				-0.012^{**}
				(0.004)
Num. obs.	508916	508916	508916	508916
R^2	0.871	0.871	0.871	0.871
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Standard errors cluster corrected at Zip Code-year level. *chinaSell* is an indicator for a Chinese seller, and *chinaBuy* is an indicator for a Chinese buyer. Individuals are classified as either Chinese or non-Chinese using the logit classifier in Equation 1. *any*4 is an indicator for the presence of any *4* in the address. *total*4 is the total number of *4*s in the address. *building*4 is an indicator for the presence of a *4* in the house number. *buildingLast*4 is an indicator if the house number ends in a *4*.

Table 8: Ethnicity and *8*s in Addresses, Multinomial Classifier

	Model 1	Model 2	Model 3	Model 4
chinaSell	-0.058^{***}	-0.056^{***}	-0.056^{***}	-0.055^{***}
	(0.003)	(0.003)	(0.003)	(0.003)
chinaBuy	-0.035^{***}	-0.035^{***}	-0.033^{***}	-0.029^{***}
	(0.003)	(0.003)	(0.003)	(0.002)
any8	0.001			
	(0.001)			
any8 × chinaSell	0.014^{**}			
	(0.005)			
any8 × chinaBuy	0.020^{***}			
	(0.003)			
total8		0.000		
		(0.001)		
total8 × chinaSell		0.008		
		(0.004)		
total8 × chinaBuy		0.015^{***}		
		(0.003)		
buildingAny8			0.001	
			(0.001)	
buildingAny8 × chinaSell			0.011	
			(0.006)	
buildingAny8 × chinaBuy			0.017^{***}	
			(0.004)	
buildingLast8				0.002^{*}
				(0.001)
buildingLast8 × chinaSell				0.019^{*}
				(0.008)
buildingLast8 × chinaBuy				0.007
				(0.005)
Num. obs.	508916	508916	508916	508916
R^2	0.871	0.871	0.871	0.871
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Standard errors cluster corrected at Zip Code-year level. *chinaSell* is an indicator for a Chinese seller, and *chinaBuy* is an indicator for a Chinese buyer. Individuals are classified as either Chinese, Korean or non-Chinese using the multinomial classifier in Equation 3. *any*8 is an indicator for the presence of any 8 in the address. *total*8 is the total number of 8s in the address. *building*8 is an indicator for the presence of an 8 in the house number. *buildingLast*8 is an indicator if the house number ends in an 8.

Table 9: Koreans and *8*s in Addresses, Multinomial Classifier

	Model 1	Model 2	Model 3	Model 4
koreaSell	-0.029^{***}	-0.029^{***}	-0.030^{***}	-0.029^{***}
	(0.004)	(0.004)	(0.004)	(0.004)
koreaBuy	0.007	0.006	0.006	0.008^{*}
	(0.004)	(0.004)	(0.004)	(0.003)
any8	0.001			
	(0.001)			
any8 × koreaSell	0.001			
	(0.007)			
any8 × koreaBuy	0.007			
	(0.006)			
total8		0.001		
		(0.001)		
total8 × koreaSell		−0.000		
		(0.006)		
total8 × koreaBuy		0.008		
		(0.005)		
buildingAny8			0.001	
			(0.001)	
buildingAny8 × koreaSell			0.005	
			(0.008)	
buildingAny8 × koreaBuy			0.012	
			(0.007)	
buildingLast8				0.003^{**}
				(0.001)
buildingLast8 × koreaSell				0.003
				(0.013)
buildingLast8 × koreaBuy				0.010
				(0.010)
Num. obs.	508916	508916	508916	508916
R^2	0.871	0.871	0.871	0.871
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Standard errors cluster corrected at Zip Code-year level. *koreaSell* is an indicator for a Chinese seller, and *koreaBuy* is an indicator for a Korean buyer. Individuals are classified as either Chinese, Korean or non-Chinese using the multinomial classifier in Equation 3. *any*8 is an indicator for the presence of any *8* in the address. *total*8 is the total number of *8*s in the address. *building*8 is an indicator for the presence of an *8* in the house number. *buildingLast*8 is an indicator if the house number ends in an *8*.

Table 10: Olympic Athlete Names and 10 Largest Multinomial Coefficients

China	$\widehat{\phi^*}$	Korea	$\widehat{\phi^*}$	United States	$\widehat{\phi^*}$
li	7.932	cho	8.153	white	5.497
liu	7.745	yeong	7.520	kevin	4.397
xu	7.451	cheol	7.469	amy	3.381
zhou	7.305	yeol	7.093	david	2.721
zhu	7.289	seong	6.781	mike	2.593
zhao	7.226	seung	6.732	craig	2.503
fai	7.172	ja	6.393	bill	2.500
zhang	7.129	shin	6.228	jim	2.466
he	7.104	gi	6.212	hugh	2.460
xie	7.100	gil	6.194	rick	2.447

Table 10 shows the 10 largest estimated regression coefficients associated with Chinese (including Mainland China, Hong Kong and Taiwan), Korean, and American names from the multinomial classifier.

Table 11: Multinomial Classifier Transaction Counts

Ethnicity Indicator	Number of Transactions
chinaSell	10,677
chinaBuy	24,869
koreaSell	2,373
koreaBuy	3,777

The ethnicity indicator variables *chinaBuy* and *chinaSell*, *koreaBuy*, and *koreaSell* are created using the multinomial classifier. "China" includes Mainland China, Hong Kong and Taiwan.

Table 12: Chinese (including Mainland China, Hong Kong and Taiwan) and *8*s in Addresses, Multinomial Classifier

	Model 1	Model 2	Model 3	Model 4
chinaSell	-0.052^{***}	-0.050^{***}	-0.050^{***}	-0.049^{***}
	(0.003)	(0.003)	(0.003)	(0.003)
chinaBuy	-0.032^{***}	-0.032^{***}	-0.030^{***}	-0.027^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
any8	0.001			
	(0.001)			
any8 × chinaSell	0.014^{**}			
	(0.005)			
any8 × chinaBuy	0.016^{***}			
	(0.003)			
total8		0.000		
		(0.001)		
total8 × chinaSell		0.008^{*}		
		(0.004)		
total8 × chinaBuy		0.012^{***}		
		(0.002)		
buildingAny8			0.001	
			(0.001)	
buildingAny8 × chinaSell			0.011^{*}	
			(0.005)	
buildingAny8 × chinaBuy			0.014^{***}	
			(0.003)	
buildingLast8				0.002^{*}
				(0.001)
buildingLast8 × chinaSell				0.011
				(0.007)
buildingLast8 × chinaBuy				0.004
				(0.004)
Num. obs.	508916	508916	508916	508916
R^2	0.871	0.871	0.871	0.871
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 13: Koreans and 8s in Addresses, Multinomial Classifier

	Model 1	Model 2	Model 3	Model 4
koreaSell	-0.025^{***}	-0.025^{***}	-0.025^{***}	-0.025^{***}
	(0.005)	(0.005)	(0.004)	(0.004)
koreaBuy	0.010^{*}	0.010^{*}	0.010^{**}	0.013^{***}
	(0.004)	(0.004)	(0.004)	(0.003)
any8	0.001			
	(0.001)			
any8 × koreaSell	−0.001			
	(0.007)			
any8 × koreaBuy	0.009			
	(0.006)			
total8		0.001		
		(0.001)		
total8 × koreaSell		−0.001		
		(0.006)		
total8 × koreaBuy		0.008		
		(0.005)		
buildingAny8			0.001	
			(0.001)	
buildingAny8 × koreaSell			0.000	
			(0.008)	
buildingAny8 × koreaBuy			0.012	
			(0.007)	
buildingLast8				0.003^{**}
				(0.001)
buildingLast8 × koreaSell				0.003
				(0.014)
buildingLast8 × koreaBuy				0.006
				(0.011)
Num. obs.	508916	508916	508916	508916
R^2	0.870	0.870	0.870	0.870
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 14: Olympic Athlete Names and 10 Largest Multinomial Coefficients

Mainland China	$\widehat{\phi^*}$	Hong Kong & Taiwan	$\widehat{\phi^*}$	United States	$\widehat{\phi^*}$
zhou	7.071	lau	8.374	amy	4.031
zhang	7.067	rull	8.253	hugh	3.161
zhu	7.060	chan	8.031	kevin	2.506
zhao	7.013	ng	7.529	mike	2.210
guo	6.847	law	7.434	david	2.116
xie	6.803	fung	7.390	craig	2.097
cai	6.590	kwok	7.305	bill	2.095
zheng	6.519	lam	7.243	jim	2.067
hou	6.512	kam	7.214	tom	2.012
zou	6.435	wong	7.160	rick	2.007

Table 14 shows the 10 largest estimated regression coefficients associated with Mainland Chinese, Hong Kong and Taiwan, and American names from the multinomial classifier.

Table 15: Multinomial Classifier Transaction Counts

Ethnicity Indicator	Number of Transactions
chinaSell	4,062
chinaBuy	11,927
hktwSell	6,106
hktwBuy	10,788

The ethnicity indicator variables *chinaBuy* and *chinaSell*, *hktwBuy*, and *hktwSell* are created using the multinomial classifier. "China" stands for Mainland China, and "hktw" stands for Hong Kong (HK) and Taiwan (TW).

Table 16: Mainland Chinese and *8*s in Addresses, Multinomial Classifier

	Model 1	Model 2	Model 3	Model 4
chinaSell	-0.066^{***}	-0.064^{***}	-0.064^{***}	-0.063^{***}
	(0.004)	(0.004)	(0.004)	(0.004)
chinaBuy	-0.042^{***}	-0.042^{***}	-0.039^{***}	-0.036^{***}
	(0.003)	(0.003)	(0.003)	(0.003)
any8	0.001			
	(0.001)			
any8 × chinaSell	0.012			
	(0.007)			
any8 × chinaBuy	0.019^{***}			
	(0.004)			
total8		0.001		
		(0.001)		
total8 × chinaSell		0.006		
		(0.005)		
total8 × chinaBuy		0.015^{***}		
		(0.003)		
buildingAny8			0.001	
			(0.001)	
buildingAny8 × chinaSell			0.009	
			(0.008)	
buildingAny8 × chinaBuy			0.015^{***}	
			(0.004)	
buildingLast8				0.003^{*}
				(0.001)
buildingLast8 × chinaSell				0.018
				(0.010)
buildingLast8 × chinaBuy				0.005
				(0.006)
Num. obs.	508916	508916	508916	508916
R^2	0.871	0.871	0.871	0.871
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 17: Hong Kong and Taiwan and *8*s in Addresses, Multinomial Classifier

	Model 1	Model 2	Model 3	Model 4
hktwSell	-0.034^{***}	-0.033^{***}	-0.033^{***}	-0.030^{***}
	(0.003)	(0.003)	(0.003)	(0.003)
hktwBuy	-0.013^{***}	-0.013^{***}	-0.012^{***}	-0.010^{***}
	(0.003)	(0.003)	(0.003)	(0.002)
any8	0.001			
	(0.001)			
any8 × hktwSell	0.012^{*}			
	(0.006)			
any8 × hktwBuy	0.011^{*}			
	(0.004)			
total8		0.001		
		(0.001)		
total8 × hktwSell		0.009		
		(0.004)		
total8 × hktwBuy		0.010^{**}		
		(0.004)		
buildingAny8			0.001	
			(0.001)	
buildingAny8 × hktwSell			0.012	
			(0.006)	
buildingAny8 × hktwBuy			0.012^{*}	
			(0.005)	
buildingLast8				0.003^{*}
				(0.001)
buildingLast8 × hktwSell				−0.000
				(0.009)
buildingLast8 × hktwBuy				0.010
				(0.007)
Num. obs.	508916	508916	508916	508916
R^2	0.870	0.870	0.870	0.870
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

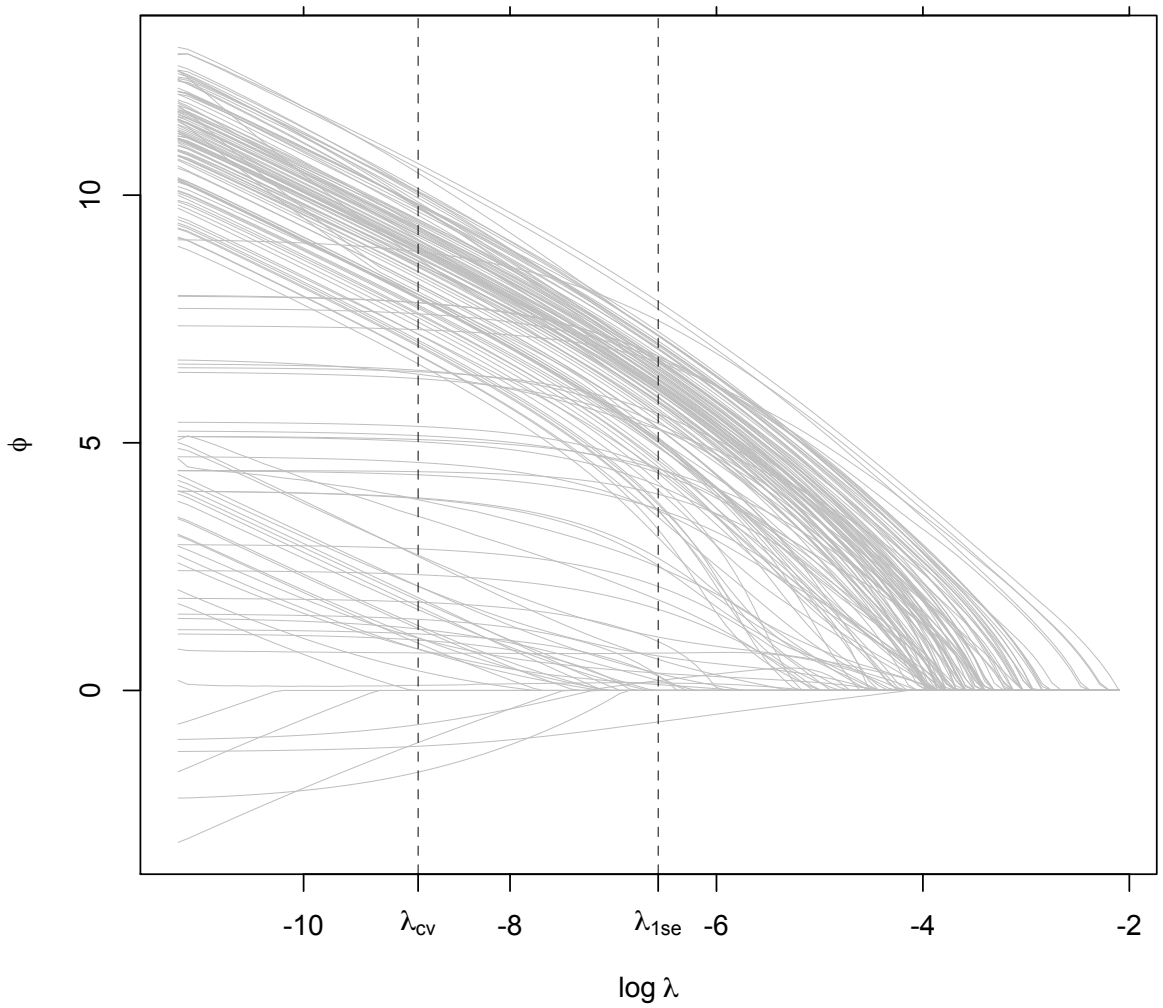
Figure 1: Olympic Athlete Names

Chinese National Team

United States National Team

100 most frequent names appearing on the Summer Olympic Games rosters for each country. More frequent names are indicated with a larger font.

Figure 2: Solution Path for $\phi^*(\lambda)$



(a) Figure 2 displays $\phi^*(\lambda)$ for various λ. $\phi^*(\lambda)$ is the solution to Equation 2. As $\lambda \to 0$, the penalty on the coefficients decreases, and $\phi^*(\lambda)$ becomes the maximum-likelihood estimator. Because of the shape of the ℓ_1 penalty, $\phi_p^*(\lambda) = 0$ for some p. λ_{cv} is the 5-fold cross-validated and λ_{1se} is the largest λ such that the cross-validated log-likelihood is within 1 standard error of the cross-validated log-likelihood when evaluated at λ_{cv}.

Figure 3: Chinese Buyers and Sellers Over Time

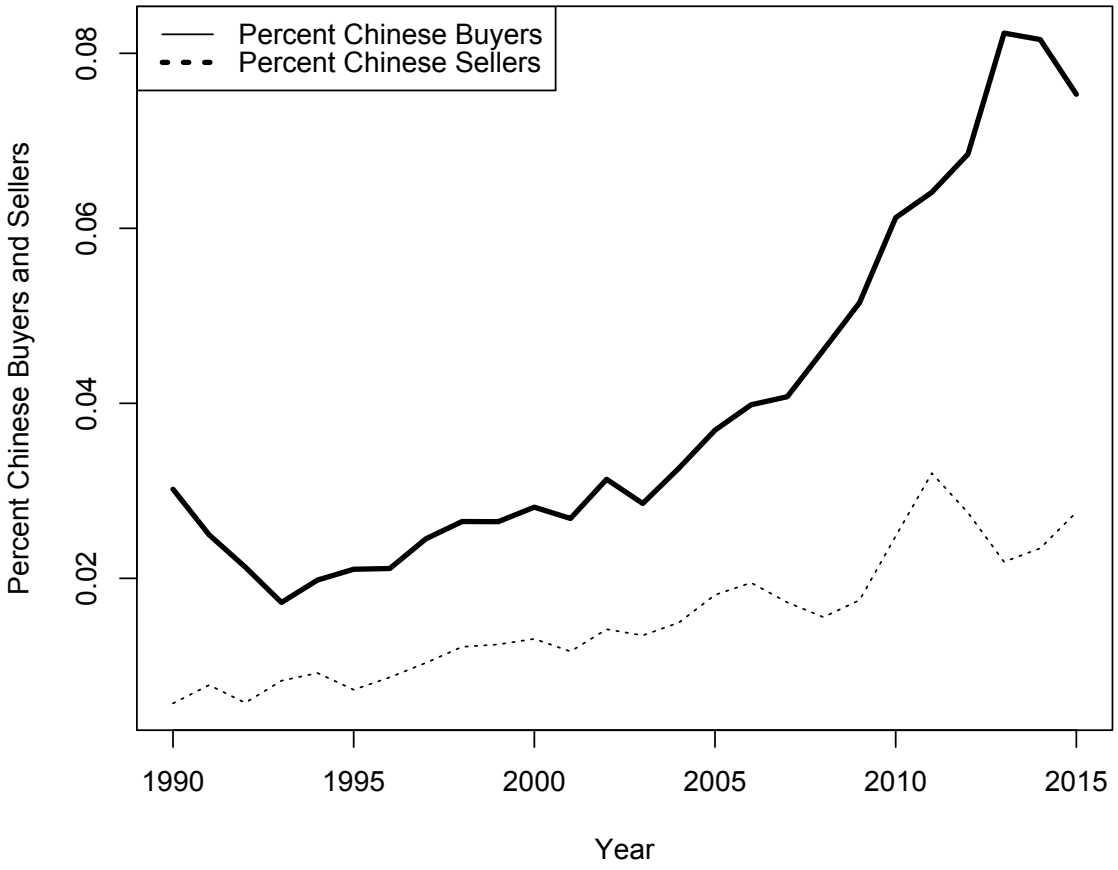


Figure 3 displays the number of Chinese Buyers and Chinese Sellers as a percentage of total transactions over time.

Figure 4: Fraction of Chinese Single Family Home Buyers by Census Tract

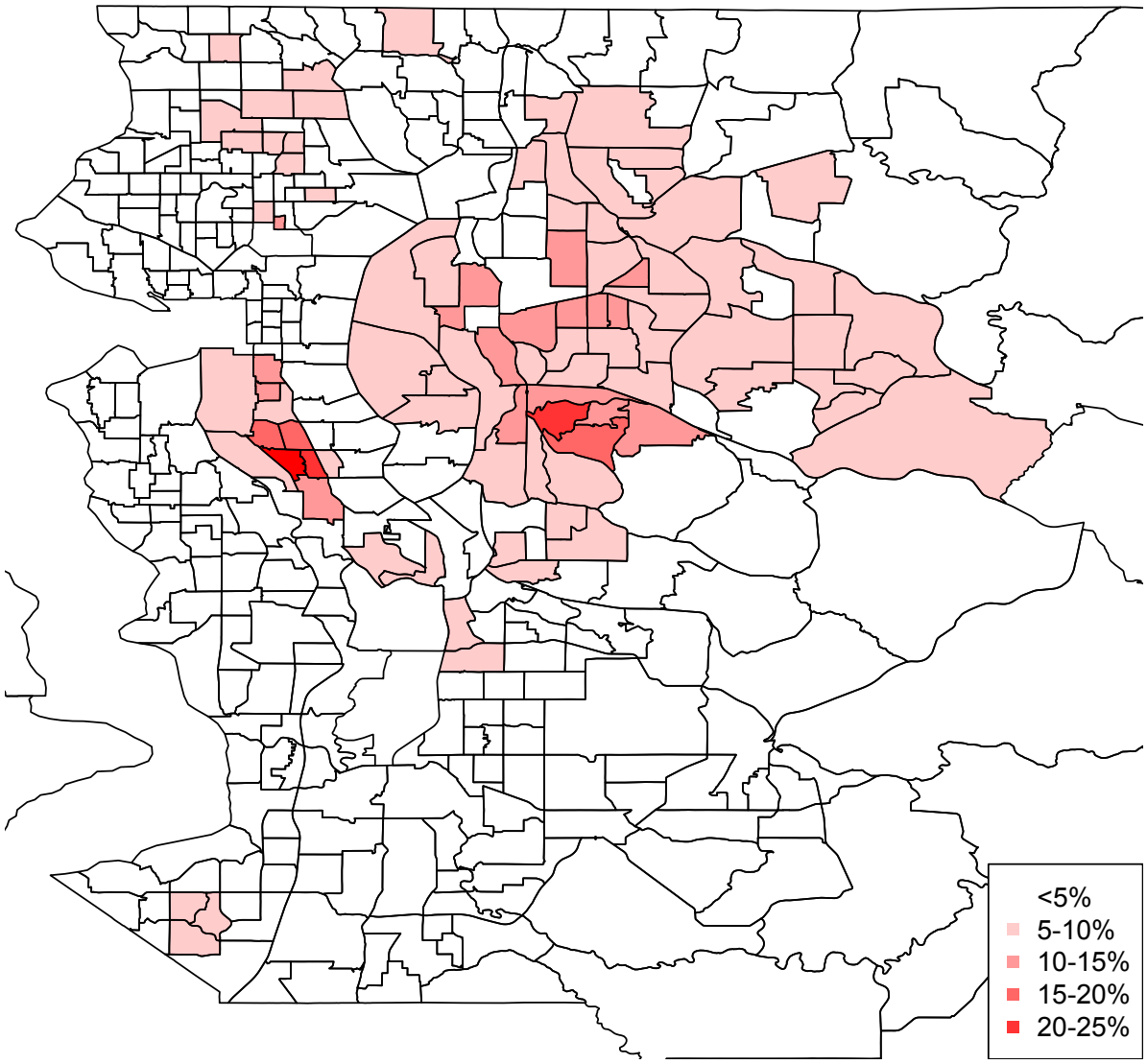


Figure 4 shows the number of Chinese single family home buyers in a given census tract as a percentage of total single family home transactions in the census tract. Total transactions begin January 1990 and end December 2015.

Figure 5: Location of Single Family Homes Purchased by Chinese Buyers

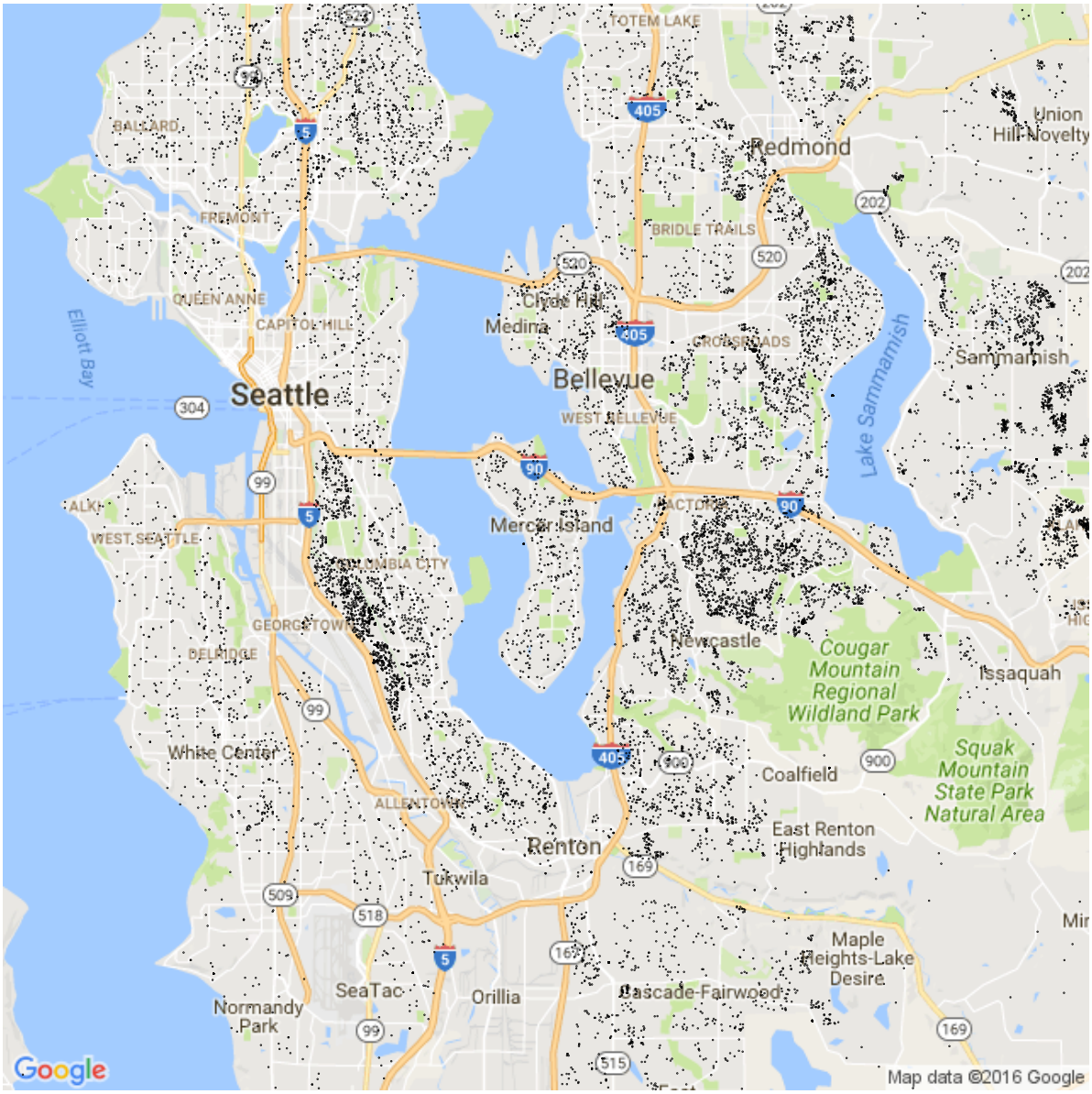


Figure 5 identifies the locations of single family homes bought by an individual identified as Chinese in Seattle over the period January 1990 to December 2015.

A Appendix: Coefficients Estimated by Bootstrapping

Murphy and Topel (1985) suggest that general two-step estimation procedures fail to account for the fact that imputed regressors are measured with sampling error. Even though the sample size is large, the coefficients estimated based on generated values may still be biased. To address this issue, we bootstrap both the Olympic roster classifier and the Seattle housing transactions regression model and estimate the coefficients 500 times. We further calculate the means and standard errors of the 500 bootstrapped results.

Table 18 displays the results estimated by bootstrapping 500 times. The results for Chinese buyers' and sellers' preferences of the "lucky" number 8 are still very comparable to the corresponding ones in Table 6 of the main manuscript.

Table 18: Hong Kong and Taiwan and *8*s in Addresses, Multinomial Classifier

	Model 1	Model 2	Model 3	Model 4
chinaSell	-0.053^{***}	-0.052^{***}	-0.051^{***}	-0.049^{***}
	(0.003)	(0.003)	(0.003)	(0.003)
chinaBuy	-0.045^{***}	-0.044^{***}	-0.042^{***}	-0.038^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
any8	0.002^{*}			
	(0.001)			
any8 × chinaSell	0.016^{**}			
	(0.006)			
any8 × chinaBuy	0.024^{***}			
	(0.004)			
total8		0.001∗		
		(0.001)		
total8 × chinaSell		0.011^{*}		
		(0.005)		
total8 × chinaBuy		0.017^{***}		
		(0.003)		
buildingAny8			0.001	
			(0.001)	
buildingAny8 × chinaSell			0.014∗	
			(0.006)	
buildingAny8 × chinaBuy			0.021^{***}	
			(0.005)	
buildingLast8				0.001
				(0.001)
buildingLast8 × chinaSell				0.011
				(0.009)
buildingLast8 × chinaBuy				0.018^{***}
				(0.005)
Num. obs.	508916	508916	508916	508916
R^2	0.815	0.815	0.815	0.815
Zip Code - Year FE	Y	Y	Y	Y

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Standard errors cluster corrected at Zip Code-year level. *chinaSell* is an indicator for a Chinese seller, and *chinaBuy* is an indicator for a Chinese buyer. Individuals are classified as either Chinese or non-Chinese using the logit classifier in Equation 1. *any*8 is an indicator for the presence of any *8* in the address. *total*8 is the total number of *8*s in the address. *building*8 is an indicator for the presence of an *8* in the house number. *buildingLast*8 is an indicator for house numbers ending in an *8*.