# Of Men and Rice: The Effect of Migrant Diversity on Households' Food Expenditure

Lin Rongxian Timothy

August 12, 2017

**Abstract**

This paper examines the effect of migrant diversity on households' expenditure on food products. Using product attribute descriptions from the Nielsen Consumer Panel dataset and mapping it to a corpus of recipes, I calculate a region weighted expenditure share for each household. Exploiting variation in migrant settlement patterns across counties and country of origin, I find that a one percentage point increase in foreign-born share of a particular region is associated with a 0.28 percentage point increase in expenditure share on consumer packaged goods associated with that region. However, a negative relationship is observed among Asian countries. The findings are robust to various means of constructing the dataset, expenditure shares and choice of instruments.

The effect of migration on the labour market outcomes of natives have been widely debated in the economics literature (Card, 1990; Borjas, 2003; Kerr and Kerr, 2011). However, fewer studies have been done on the broader multidimensional effects of migration, such as the transmission of culture. Existing work on cultural diversity tends to focus on its economic value evaluated through labour market outcomes (Ottaviano and Peri, 2006), housing prices (Bellini et al., 2013) or innovation (Kerr and Lincoln, 2010; Hunt and Gauthier-Loiselle, 2010), but there has not been any empirical studies documenting the transmission of migrant culture. The diversity of cuisine styles and the variety ingredients used provide an opportunity to test such effects. Furthermore, the popularity of different food options in cities worldwide suggests that if there were true spillover effects of migrant culture, preferences over food would be an obvious candidate.

This study analyses the effect of migration (using foreign-born share as a proxy) on food consumption patterns as evidenced from households' expenditure on consumer packaged goods. Using a novel corpus of food related terms extracted from recipes and cookbooks' indexes, and drawing on text classification methods, I construct an index measuring the association of each term to a particular country or region. Merging the term-region index with product attributes from households' expenditure records from the Nielsen Homescan Consumer Panel allows me to calculate a region weighted expenditure share for each household. Subsequently, I exploit the variation of migrant settlement patterns across counties and country of origin to analyse the effect of foreign-born share on consumption expenditure.

I find strong evidence that a county's foreign-born share affects present day consumption patterns of natives. Using lagged foreign-born share as an instrument, I find that a percentage point increase in foreign-born share from a particular region is associated with a 0.28 percentage point increase in expenditure share on food products associated with that particular region. However, a negative relationship between foreign-born share and expenditure is observed when the sample is restricted to only Asian countries. A percentage point increase in foreign-born share from Asian countries is associated with a 0.13 percentage point decrease in expenditure share. These findings hold across different product groups and are robust to alternative means of constructing expenditure shares and choices of datasets.

The construction of the term-region index is similar to existing works in the literature which tap on text classification methods to uncover interesting economic relationships (Antweiler and Frank, 2004; Gentzkow and Shapiro, 2010; Baker et al., 2016). This paper adapts the term frequency-inverse document frequency (TF-IDF) model, a popular approach in text classification problems, and applies it to a new context — household consumption. While the mapping from recipe ingredients to consumer packaged goods may be somewhat arbitrary, I show that the constructed index accords to intuition and that the results are robust to alternative means of construction.

While past studies examining the link between migration and trade in goods and services (Gould, 1994; Rauch and Trindade, 2002), tend to find a positive relationship between both variables, it is hard to infer from cross-country data who the agents of consumption are. This paper provides a direct

look at the consumption behaviour of natives by tapping on household level expenditure data. More broadly, the study is also related to the literature examining the formation of consumers' preferences. Extensive studies in food tastes have shown that social and environmental factors play a large part in shaping one's food preferences (Rozin and Vollmecke, 1986; Nestle et al., 1998; Birch, 1999). Closer to the industrial organisation literature, Bronnenberg et al. (2012) show that the preferences of interstate migrants within the US over consumer packaged goods converge slowly to native preferences. Could the effect be true the other way around? Exploiting variation in the settlement patterns of foreign-born migrants to explain present day consumption behaviour provides an opportunity to understand how taste preferences are shaped by one's living environment.[1]

The paper is organised as follows. Section 1 introduces the data. Section 2 documents the construction of the term-region index used in the calculation of expenditure shares. Section 3 explains the empirical methodology and estimation strategy. Section 4 presents the main results and section 5 concludes.

# 1 Data

## 1.1 Households' Expenditure and Characteristics

Expenditure data at the household level is obtained from the 2011 Nielsen Homescan Consumer Panel dataset. The panel is made up of 62,092 households

[1]Such spillover effects are connected to the broader literature on peer effects and economic networks (Manski, 1993; Brock and Durlauf, 2001; Jackson, 2010).

drawn from 2708 counties across 49 states. Each household uses in-home scanners to record their purchases. Household purchase records are available at the universal product code (UPC) level and covers both food and non-food items across all U.S. retail outlets. This is supplemented with additional information on when and where the purchase was made, the amount spent and attributes of the product.[2] Background information on the panellist containing details on their race, age, education and income levels are also available.

I restrict the panel to white, non-hispanic households and examined purchases across the following eight food related product groups: condiments, gravies and sauces; vegetables-canned; seafood-canned; prepared food (ready to serve); prepared food (dry mixes); prepared food (frozen); pasta; spices, seasoning and extracts.[3] Each product at the UPC level contains additional information on its attributes including information on the product's brand and a brief product description.[4] I map each product description to a term-region index which assigns a weight to a term-region pair. More details on the construction of the index is presented in the subsequent section. The final dataset contains 62,729 unique UPC codes.

Region weighted expenditure is calculated using the actual price paid less any discounts from coupons. Expenditure shares are computed based on all the goods purchased across the eight product groups for the entire year. All

[2]The level of detail in the dataset makes it a popular choice in the industrial organisation and marketing literature. Einav et al. (2008) provides a validation study of the dataset.

[3]While it would be ideal to track expenditure on fresh produce, such expenditure is grouped under the broad "magnet" category and it is not possible to distinguish what produce the household is actually purchasing.

[4]Examples of product descriptions include, "beans", "navy beans", "mahi mahi mango marinaded" and "fajita seasoning".

results presented in the subsequent sections are weighted by the projection factor provided in the dataset.

## 1.2 Corpus of Food Related Terms

6275 recipes were scrapped from the "World Cuisine" section of allrecipe.com, a food-centric website which contains recipes submitted by community members. Recipes in the "World Cuisine" section are tagged to a particular country or region which I further aggregated into twenty-four different regions.[5] The corpus of food related terms comprises of the recipes' title as well as the lists of ingredients and is used to construct the term-region index.

As an additional data source and to validate the information obtained from the recipe dataset, I also compiled information from indexes of recipe books. 161 books from the Vancouver Public Library, selected following the same categories of the recipe dataset, were scanned and digitalised. While the recipe dataset provides an indication of how frequently an ingredient is used for a particular regional cuisine, the book indexes dataset gives a more authoritative take on how important a particular recipe is to the cuisine.

## 1.3 Foreign-born and county level information

County level data on shares of foreign-born and other geographical information are obtained from the 2010 and 1980 NHGIS dataset. Foreign-born shares are

[5]They are: Oceania, Other West and Central Europe, North America, Caribbean, China, Africa, Eastern Europe, Philippines, France, Germany, Greece, India, Other Southeast Asia, Middle East, Italy, Japan, Korea, Central America, Scandinavia, South America, Spain, Thailand, Uk and Ireland, and Vietnam.

aggregated to match the regions used in the term-region index. Due to data availability limitations of the 1980 NHGIS, some countries have been aggregated to larger regions.[6] Table B1 in Appendix B lists the regions included in the regression analysis.

# 2 Construction of the Term-region Index and Expenditure Shares

## 2.1 Rationale

A key component of this paper is the construction of a region weighted expenditure share based on household purchases detailed at the UPC level. One approach would be to use the country of origin for each product, though doing so would ignore the fact that products imported from a particular country may not actually be related to it. The Nielsen dataset also does not contain information on a product's country of origin. A second alternative would be to use the brands of the products as a signal. While information on brand ownership is available it does not seem to be a good way to derive a mapping as ownership is often very different from the brand marketing. A third option would be to search for the mention of a particular country in the product description of the dataset. However, such a straightforward matching strategy would severely limit the number of matched products in the analysis since the majority of products do not have country names in the description field.

[6]The 1970 IPUMS, an alternative data source containing more detailed foreign-born data but over fewer counties, is also used as a robustness check.

The approach implemented in this paper uses product descriptions as a signal of how closely related it is to a particular region. It could be regarded as a fuzzy matching variant of the third approach and is built on the idea that certain types of food tend to be consume in some regions more so than others.[7] Two characteristics stand out with such an approach. First, the matching is fuzzy - there may not be a 100% match for a particular region and product. Instead, each product-region pair is assigned a weight that corresponds to the probability that the product is correctly associated with that region. Second, a separate corpus of information is needed to serve as the knowledge base so as to map each product term to a term-region weight.

## 2.2 Methodology

The term-region index maps product attributes to term-region weights which are used to calculate product-region expenditure shares. The index is constructed using a "bag-of-words" model, a commonly used algorithm in Natural Language Processing which treats text as a bag (multiset) of words.[8] It takes into account the words used and the frequency which they occur in a particular document but disregard the semantic relationship in the sentences.[9]

Geographical regions take the place of documents while ingredients take the place of words. To account for the fact that phrases might contain more

[7]For example, Basmati rice is more closely associated with Indian food, while Japanese rice tends to be more widely consumed in North-east Asia.

[8]It has been used in the field of Linguistics (Harris, 1954), and subsequently in information retrieval and classification tasks (Jones, 1972).

[9]Disregarding the semantics or grammar of the sentence is a beneficial simplification as most recipes contain only phrases of ingredients or instructions and the semantic structure does not convey additional information.

information than individual words, I also implemented a "2-gram" model, which takes into consideration word pairs.[10] For simplicity, I shall refer to each word or word-pair as a term.

To measure term-region association, I calculate a term frequency-inverse document frequency (TF-IDF) score for every term-region pair that is found in the recipe or book collection. The TF-IDF formula is given by:

$$TF\text{-}IDF_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot log \frac{N}{n_t} \tag{1}$$

where $f_{t,d}$ is the frequency in which term $t$ appears in document $d$, $N$ is the total number of documents and $n_t$ is the total number of documents where term $t$ is found. The first component on the right hand side is the term frequency and it measures the relative frequency which a term is used and could be regarded as a signal of how common a particular term is. This is multiplied by the inverse document frequency component which down-weights common terms used across documents.[11] The TF-IDF approach is commonly used in document classification problems and usually out performs multinomial naive bayes (Kibriya et al., 2004).

To illustrate the algorithm, Figure 1 compares the term frequencies of words across three regions, China, Japan and Central America. Words at the top right hand corner of the graph are equally used by both regions. Words

[10]For example, in the "bag-of-words" model, the term "white jasmine rice" would be analysed as three separate words: "white", "jasmine" and "rice." In the "2-gram" model, pairs of words would be considered: "white jasmine" and "jasmine rice."

[11]The most commonly used terms in the recipe dataset include words related to measurement, like "cup", "teaspoon", and "tablespoon", as well as kitchen staples, like "salt", "pepper", and "oil." The inverse document frequency term ensures that the weight of such terms are given a score of 0 since they are found across all documents.

at the top left corner appear much more frequently in the recipes tagged as "China", while those at the bottom right corner appear more frequently in recipes tagged as "Central America" or "Japan." The location of the words supports one's intuition of distinctive ingredients used in a particular cuisine. For example, "Sesame" appears much more frequently in Chinese than Central American food. The dispersion of the scatter plot also supports the idea that recipes tagged with "China" are more closely related to recipes tagged with "Japan" than "Central America." Comparing the TF-IDF scores across regions, the correlation between China and Japan is 0.83, while the correlation between China and Central America is 0.55.

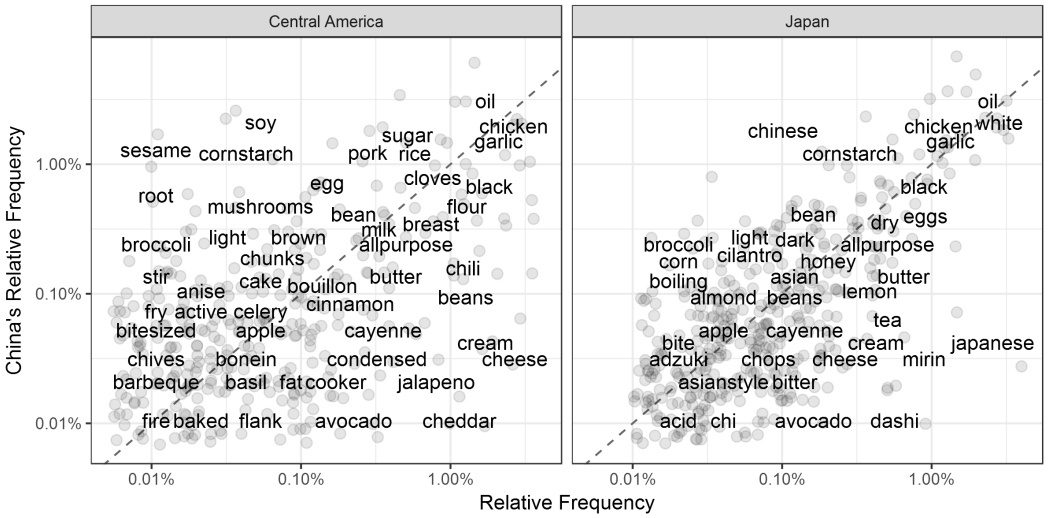


Figure 1: Comparison of Term Frequencies

A list of the top 10 words by TF-IDF score for these three regions is shown in Figure 2. Notice that words which are common across all regions such as "oil" and "chicken" are not in the list. Rather, ingredients which feature prominently in a particular cuisine but not in others tend to be the highest

scorers.

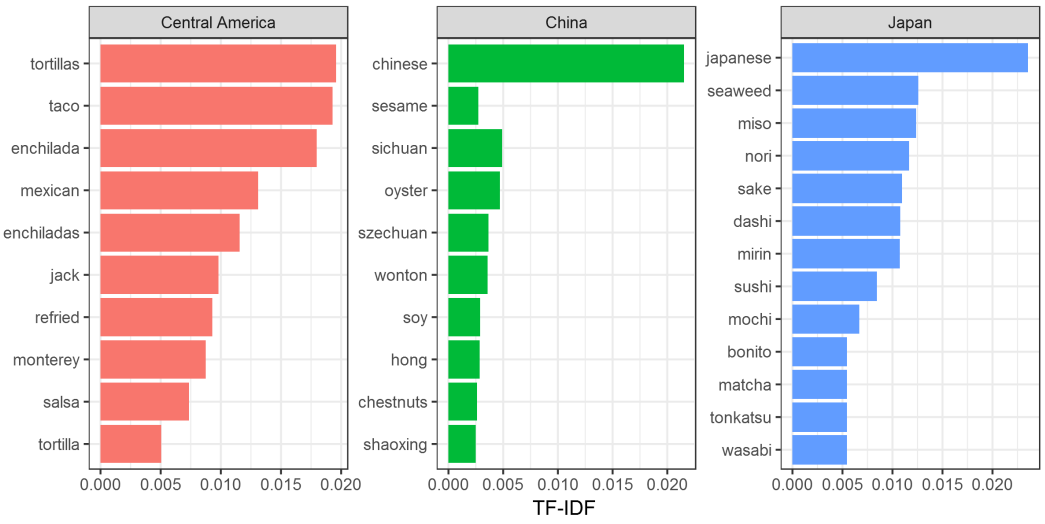


Figure 2: Top 10 TF-IDF Scores by Region

The TF-IDF scores are subsequently normalised to sum to one across all regions. A similar set of TF-IDF scores were calculated using the book indexes as input. The correlation between the two set of scores is 0.77. The final set of term-region scores were constructed by averaging across both scores and normalising them such that they to sum to one across all regions. The implicit assumption made in the construction of the term-region index is that each region corresponds to a distinct food culture and the TF-IDF scores capture the level of association between terms and regions. In reality, the transmission of food culture is not necessary restricted to national boundaries and one would expect that the ingredients used in neighbouring regions would be correlated. The regression results in Appendix C include robustness checks using more aggregated regions to ensure that each defined region is unique and the correlation between regions are lower.

This set of scores were merged with the product description field to generate product-region weights based on the following rules: Matches based on word pairs are given priority over single word matches; If there are multiple matches, the TF-IDF scores of the term with the highest maximum term-region score is chosen.[12] Examples of products by region and scores are provided in Appendix B, Table B2.

Summary statistics of the product scores are shown in Table 1. There are a large number of missing values across product-region weights. This suggests that while there are certain terms that are common across all regions, most of the matched terms are used only in specific cuisines. Certain regions such as Southeast Asia also tend to have relatively poorer match rates. Conditional on matching, the matching algorithm is most confident of its ability to identify products associated with Italy or Central America. This is seen by the relatively high mean weights of 0.36 an 0.37 respectively, as well as the relatively large number of products with weights greater than 0.5 and 0.75.

[12]For example, a product with description of "fajita seasoning" would be matched with the word pair "fajita seasoning" if available or the TF-IDF scores of "fajita" and "seasoning." In this example, the TF-IDF scores of "fajita" was used because it has a very high score for Central America compared to the term "seasoning" which is more generic.

Table 1: Summary Statistics of Product-Region Weights

| | N | Missing | >0.5 | >0.75 | Mean | SD |
|---|---|---|---|---|---|---|
| Africa | 9,541 | 53,188 | 146 | 145 | 0.09 | 0.12 |
| Caribbean | 19,168 | 43,561 | 852 | 611 | 0.13 | 0.18 |
| Central America | 19,734 | 42,995 | 6,237 | 4,184 | 0.37 | 0.38 |
| China | 14,178 | 48,551 | 806 | 580 | 0.15 | 0.20 |
| Eastern Europe | 13,873 | 48,856 | 420 | 125 | 0.16 | 0.15 |
| France | 17,185 | 45,544 | 438 | 325 | 0.10 | 0.16 |
| Germany | 7,430 | 55,299 | 556 | 122 | 0.20 | 0.17 |
| Greece | 17,073 | 45,656 | 304 | 202 | 0.16 | 0.20 |
| India | 11,516 | 51,213 | 1,481 | 1,413 | 0.22 | 0.31 |
| Italy | 21,822 | 40,907 | 7,451 | 4,371 | 0.36 | 0.37 |
| Japan | 12,711 | 50,018 | 1,654 | 1,593 | 0.23 | 0.30 |
| Korea | 7,751 | 54,978 | 331 | 303 | 0.18 | 0.18 |
| Middle East | 13,816 | 48,913 | 334 | 251 | 0.12 | 0.18 |
| Other Southeast Asia | 7,008 | 55,721 | 1,431 | 103 | 0.21 | 0.24 |
| Philippines | 13,937 | 48,792 | 527 | 147 | 0.13 | 0.15 |
| Scandinavia | 7,352 | 55,377 | 456 | 341 | 0.19 | 0.20 |
| South America | 9,347 | 53,382 | 236 | 62 | 0.12 | 0.12 |
| Spain | 9,658 | 53,071 | 387 | 46 | 0.17 | 0.16 |
| Thailand | 15,312 | 47,417 | 656 | 317 | 0.10 | 0.17 |
| Vietnam | 3,676 | 59,053 | 232 | 22 | 0.17 | 0.17 |
| Maximum Weight | 46,929 | 15,800 | 24,935 | 15,263 | 0.60 | 0.30 |

Notes: Weights are calculated by normalising the TF-IDF scores of each product such that they some to one across all regions. Weights are bounded between 0 and 1. The first two columns counts the number of observations which are non-missing and missing. The next two columns counts the number of observations which satisfy the criteria of having weights above 0.5 and 0.75 respectively. Maximum Weight refers to the highest weight of each product across all regions.

# 3 Empirical Strategy

Under the assumption that consumer preferences take the form of a Dixit and Stiglitz (1977) type utility function, an increase in migrant inflow from region $k$ is expected to result in an increase in expenditure share of food products related to region $k$. In the model outlined in Appendix A, the increase in expenditure share could be driven by a change in consumers' preferences, a decrease in cost of goods related to region $k$ or greater product variety of goods from region $k$.

As a baseline specification, I regress the expenditure share of household $i$ from county $j$ that is associated with region $k$ on county level foreign-born share in 2010:

$$C_{ijk} = \beta FB_{jk} + \rho Dist_{jk} + \gamma_j + \mu_k + \epsilon_{ijk} \tag{2}$$

where $\gamma$, $\mu$ are county and region fixed effects. $\beta$, the parameter of interest, gives the average effect of a percentage point increase in foreign-born share on the expenditure share of consumers. $Dist$ represents the log pairwise distance between county $j$ and region $k$.

The fixed effect specification alleviates potential endogeneity concerns that might arise due to unobserved differences across counties which might make a particular county more welcoming to foreigners and attract natives who prefer foreign related goods. Log pairwise distance controls for county-region variation in cost of importing which might be correlated with both foreign-born and

expenditure shares. Under the assumption that foreign-born share is uncorrelated with other county-region unobservables that might affect consumption, $\beta$ can be treated as a causal estimate of migration on local consumption.

Nonetheless, one might still be concerned about the proposed specification and other endogeneity issues such as endogenous peer effects (Angrist, 2014).[13] The household background questionnaire from the Nielsen survey does not contain a field to distinguish between foreign ancestry or migration status and it is likely that a county with higher share of foreign-born from a particular region also contains more households from such regions in the Nielsen survey. To alleviate these concerns, I restrict the Nielsen panel only to white individuals of non-hispanics origin and run the regressions using both the full sample and a subset containing only Asian countries.[14]

More generally, the specification is relatively robust to problems of omitted variable bias. While there are many potential variables that are correlated to foreign-born share, these variables have to be correlated with local food consumption patterns to pose an issue. For example, one might have concerns about assortative migration of locals and migrants where people with similar preferences and food taste select into similar cities. However, it does not seem probable that taste preferences would be a key factor in the migration decision of natives.

For additional robustness and to alleviate measurement error bias, I also

[13]For example, one may be worried that natives residing at the states to the south of the US border have Latin American roots while those living in the northern states have ties with Canada. This is correlated with their food preferences as well as contemporaneous migrant flows.

[14]The assumption made here is that white individuals probably do not have Asian ancestry.

run an instrument variable regression using the share of foreign-born from region k living in to county j in 1980 as an instrument. This builds on the fact that historical settlement patterns have been shown to be a good predictor of present day migrant concentration (Altonji and Card, 1991; Card, 2001; Ottaviano and Peri, 2006). Assuming that historical settlement patterns only affect present day consumption through current foreign-born share, one would be able to recover a causal estimate of migration on consumption. This assumption holds if there is no temporal correlation in county-region unobservable factors, other than foreign-born share, that might also affect consumption. The thirty year period between both census years helps ensure that this condition holds.[15]

# 4 Results

Table 2 shows the ordinary least squares (OLS) estimates of regressing expenditure share on foreign-born share. Each specification is evaluated on two samples, the full sample and a subset of only Asian countries. The list of regions included in the regressions discussed in this section are shown in Appendix B Table B1. All regressions are weighted by the household projection factor to obtain results that are nationally representative.

Columns 1 and 2 show a positive correlation between expenditure share and foreign-born share even after controlling for geographical regions. A one

[15]Nonetheless, there may be some cause of concern if certain factors such as the attractiveness of certain cities to migrants from particular countries hold over time. This would lead to the OLS estimates being biased upwards but the direction of the bias in the IV regression is unclear and depends on the cross-correlation between the variables.

percentage point increase in foreign-born share of a particular region is associated with a 0.16 percentage point increase in expenditure share on products related to that region. However, the coefficient of foreign-born share in the Asia sample is small and not significant. Controlling for county fixed effects, the coefficients of foreign-born share in columns 3 and 4 is positive but not significant. Columns 5 and 6 correspond to the baseline specification outlined in the empirical strategy section. Controlling for both region and county fixed effect, a percentage point increase in foreign-born share of a particular region is associated with a 0.18 percentage point increase in expenditure share. However, the coefficient in the Asia sample is negative. This means that a percentage point increase in the share of foreign-born from Asian countries is associated with a 0.05 percentage point decrease in expenditure share.[16]

Table 3 presents the results of the instrumental variable (IV) estimates. The sample used is smaller than the OLS estimates as the 1980 NHGIS survey contain less detailed information on foreign-born shares compared to the 2010 survey. Column 2 shows that the OLS regression on the smaller sample yields similar results to the baseline regression reproduced in column 1.

The first stage regressions in columns 3 and 5, which corresponds to the full and Asia sample respectively, have a F-statistic above 50. This shows that foreign-born share in 1980 is a relevant instrument for foreign-born share in 2010. Turning to the instrumental variable results in column 4 and 6, the coefficient of foreign-born share for the IV full sample is larger than the OLS

[16]The difference between the full and Asia sample is largely driven by the difference in within variation across the Asia and the non-Asia (Europe and Central America) sample. The coefficient of foreign-born share is also positive and significant if the sample is restricted to all countries excluding Asia.

Table 2: OLS Estimates

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Exp Share | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| 2010 FB Share | 0.1619*** | 0.0011 | 0.3312 | 0.0791 | 0.1764*** | −0.0490 |
| | (0.0558) | (0.0174) | (0.2501) | (0.1480) | (0.0593) | (0.0365) |
| Sample | Full | Asia | Full | Asia | Full | Asia |
| Region FE | X | X | | | X | X |
| County FE | | | X | X | X | X |
| Observations | 796,448 | 448,002 | 796,448 | 448,002 | 796,448 | 448,002 |

Notes: Two-way standard errors clustered by county and region in parentheses. Full sample consists of 18 regions, Asia sample consists of 9 regions.*p<0.1; **p<0.05; ***p<0.01.

model at 0.28 while the coefficient on foreign-born share for the IV Asia sample is now -0.13 and is significant at the 5% level.[17]

The detailed product level data also allows one to separately test for changes in consumption patterns within product groups rather than over a household's entire consumption basket.[18] This would also alleviate potential concerns on the representativeness of particular product groups since certain groups, such as pasta, are more closely associated with particular regions. I focus on four product groups which contain products across all regions: prepared food (frozen), prepared food (ready to serve), condiments, gravies and sauces and spices, seasoning and extracts.

[17]The results seems consistent with the idea that attenuation bias due to measurement error was the main issue in the OLS estimates.

[18]This would be consistent with consumers having nested CES preferences over groups at the product group level.

Table 3: IV Estimates

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Exp Share | | 2010 FB Share | Exp Share | 2010 FB Share | Exp Share |
| | OLS | OLS Subset | First Stage | IV | First Stage | IV |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| 2010 FB Share | 0.1764*** | 0.2004*** | | | | |
| | (0.0593) | (0.0426) | | | | |
| 1980 FB Share | | | 1.8307*** | | 2.1091*** | |
| | | | (0.1778) | | (0.2753) | |
| Fitted 2010 FB Share | | | | 0.2762*** | | −0.1337** |
| | | | | (0.0480) | | (0.0669) |
| First Stage F-stat | | | 105.99 | | 58.71 | |
| Sample | Full | Full | Full | Full | Asia | Asia |
| Region FE | X | X | X | X | X | X |
| County FE | X | X | X | X | X | X |
| Observations | 796,448 | 595,920 | 595,932 | 595,920 | 248,305 | 248,300 |

Notes: Two-way standard errors clustered by county and region in parentheses. OLS full sample consists of 18 regions, OLS subset and IV full sample consists of 12 regions, IV Asia sample consists of 5 regions. *p<0.1; **p<0.05; ***p<0.01.

Table 4: IV Estimates by Product Category

| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Category Exp Share | | | | | | | |
| | IV | IV | IV | IV | IV | IV | IV | IV |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Fitted 2010 FB Share | 0.3246*** | −0.0835*** | 0.5860*** | −0.2106 | 0.1438 | −0.1101** | 0.0029 | 0.0205*** |
| | (0.0559) | (0.0322) | (0.1748) | (0.2660) | (0.1364) | (0.0482) | (0.0204) | (0.0026) |
| First Stage F-stat | 110.62 | 59.85 | 115.17 | 58.91 | 107.52 | 59.44 | 101.56 | 57.44 |
| Sample | Full | Asia | Full | Asia | Full | Asia | Full | Asia |
| Product category | Frozen | Frozen | Ready-to-serve | Ready-to-serve | Sauces | Sauces | Spices | Spices |
| Region FE | X | X | X | X | X | X | X | X |
| County FE | X | X | X | X | X | X | X | X |
| Observations | 564,672 | 235,280 | 557,076 | 232,115 | 586,908 | 244,545 | 539,172 | 224,655 |

Notes: Two-way standard errors clustered by county and region in parentheses. IV full sample consists of 12 regions, IV Asia sample consists of 5 regions. Expenditure shares are calculated based on households' spending within a particular product group. *p<0.1; **p<0.05; ***p<0.01.

Table 4 shows the IV results on both the full and Asia sample with expenditure shares defined over individual product groups. Across all four product groups the coefficient on the full sample remains positive, though for the condiments, gravies and sauces and spices category the effect is effectively zero. On the other hand, the coefficient on the Asia subset is negative across all but one product group. While it is positive and significant for the condiments, gravies and sauces and spices category, the coefficient of 0.02 suggests that the actual effect is very small.

While the expenditure share measure in the previous regressions is calculated based on a weighted average approach, one could also consider other methods of construction. Two other methods, a majority approach (where the region with the highest score is allocated the entire expenditure value) and a large majority approach (where the list products are restricted to those with weights greater than 0.5) are implemented in the estimates shown in Table 5. The latter approach ensures that only products which are closely related to a particular region are included in the calculations. The results are similar across all three methods with the coefficient on foreign-born share ranging from 0.28 to 0.29 in the full sample and -0.15 to -0.12 in the Asia sample.

Appendix C shows that the results are robust to using core based statistical area (CBSA) instead of counties as well as various geographical region groupings and choices of dataset for the instrument. The results pose an interesting puzzle. While the positive coefficient on the full sample accords with prior intuition, it could also be a result of the peer effect bias highlighted in the previous section. The negative relationship in the Asia sample also raises que-

stions on the exact mechanism driving this result. One possibility is that the share of foreign-born from Asia is also positively correlated with the number of restaurants selling Asian food.[19] With the availability and affordability of such options, consumers may substitute away from cooking Asian related cuisines. Alternatively, consumers living in areas with higher share of foreign-born from Asia may demand higher quality Asian food and find existing supermarkets' ready to eat options or frozen produce unsatisfactory.

Table 5: IV Estimates Using Alternative Expenditure Measures

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Exp Share | | Exp Share | | Exp Share | |
| | IV | IV | IV | IV | IV | IV |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Fitted 2010 FB Share | 0.2762*** | −0.1337** | 0.2854*** | −0.1165 | 0.2818*** | −0.1457** |
| | (0.0480) | (0.0669) | (0.0484) | (0.0884) | (0.0527) | (0.0593) |
| First Stage F-stat | 105.99 | 58.71 | 105.99 | 58.71 | 105.99 | 58.71 |
| Sample | Full | Asia | Full | Asia | Full | Asia |
| Expenditure measure | Wtd Avg | Wtd Avg | Majority | Majority | Large Majority | Large Majority |
| Region FE | X | X | X | X | X | X |
| County FE | X | X | X | X | X | X |
| Observations | 595,920 | 248,300 | 595,920 | 248,300 | 595,920 | 248,300 |

Notes: Two-way standard errors clustered by county and region in parentheses. IV full sample consists of 12 regions, IV Asia sample consists of 5 regions. Expenditure measure constructed using a weighted average approach (Wtd Avg) or with weight=1 for the region with highest weight (Majority) or with weight=1 for the region with highest weight conditional on the initial weight being greater than 0.5 (Large Majority). *p<0.1; **p<0.05; ***p<0.01.

[19]This implies that the exclusion restriction assumption fails to hold.

# 5 Conclusion

This paper analyses the effect of migration diversity on households' expenditure on consumer packaged goods. Using a novel mapping of term-region scores constructed from recipes and cookbooks' indexes, I derive a measure of a household's region-weighted expenditure share. On average, a percentage point increase in foreign-born share associated with a particular region leads to a 0.28 percentage point increase in households' expenditure share on products related to that region. However, this relationship is negative when the sample is restricted to only Asian countries.

These findings are robust to various ways of constructing the expenditure measure and hold even within the different product groups. Future research could try to uncover the mechanisms that drive these results. Supplementing the existing dataset with expenditure on fresh produce may provide a more holistic picture of household consumption patterns. In addition, examining the link between foreign-born share, restaurant availability and consumption may be another promising avenue to explore.

# A Appendix: Model

In this section, I describe a simple model that captures the effect of migration on consumption pattern of natives. Consumer preferences take the form of a Dixit-Stiglitz type utility function:

$$U_i = C_i^{\alpha_i} F_i^{1-\alpha_i} \tag{3}$$

$$F_i = \left( \sum_{j=1}^{N} \beta_j^{\frac{1}{\sigma}} x_j^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \tag{4}$$

where $C$, represents general consumption while $F$, represents the total consumption of food products and is equivalent to the sum over all variety of products $x_j$. Due to the Cobb-Douglas preference structure, one can just focus on the demand and consumption shares within food products. Consumers maximise their utility, $F_i$, subject to a total food expenditure constraint of $m$. The consumption of product $j$ relative to $k$ is given by:

$$\frac{x_j}{x_k} = \frac{\beta_j}{\beta_k} \left( \frac{p_k}{p_j} \right)^{\sigma} \tag{5}$$

Summing over all products we obtain:

$$m = \sum_{j=1}^{N} p_j x_j = \left( \sum_{j=1}^{N} \frac{\beta_j}{\beta_k} p_j^{1-\sigma} \right) p_k^{\sigma} x_k \tag{6}$$

The Marshallian demand is given by:

$$x_k = \frac{m p_k^{-\sigma}}{\sum_{j=1}^{N} \frac{\beta_j}{\beta_k} p_j^{1-\sigma}} \tag{7}$$

$$= \frac{\beta_k m}{p_k^{\sigma}} \left( \sum_{j=1}^{N} \beta_j p_j^{1-\sigma} \right)^{-1} \tag{8}$$

$$= \frac{\beta_k m}{P} \left( \frac{P}{p_k} \right)^{\sigma} \tag{9}$$

where $P = \left( \sum_{j=1}^{N} \beta_j p_j^{1-\sigma} \right)^{\frac{1}{1-\sigma}}$ and could be regarded as the general price index over all food products.

Thus expenditure share is given by:

$$s_k = \beta_k \left( \frac{P}{p_k} \right)^{\sigma-1} \tag{10}$$

Holding price constant, a change in preference modelled as an increase in a specific $\beta$ parameter results in an increase in expenditure share of that product. This model assumes that expenditure share is independent of income. While this might seem unrealistic, the use of expenditure shares defined on smaller subsets of food product categories makes this assumption more plausible.

On the production side, I assume a monopolistic competition setting where single-product firms engage in bertrand price competition and aim to maximise profit by deciding on the type of products to stock and sell. While a more realistic model would consider a multi-product firm, the simplified model sufficiently captures the idea of expanding product variety.[20] Assuming

[20] A multi-product firm such as a supermarket would base its decision to import a new product taking into account possible cannibalisation effect on sales of other products. Price

that all firms are identical, symmetric and produce only one product, the profit function for a firm importing product $i$ is given by:

$$\pi_i = p_i x_i - c_i x_i - F_i \tag{11}$$

where $c_i$ is the variable cost and $F_i$ is the fixed cost involved in producing or importing product $x_i$. Taking the first order condition with respect to price we obtain:

$$0 = x_i + p_i \frac{\partial x_i}{\partial p_i} - c_i \frac{\partial x_i}{\partial p_i} \tag{12}$$

$$c_i = \left(1 + \frac{1}{\varepsilon_d}\right) p_i \tag{13}$$

$$= \left(1 - \frac{1}{\sigma}\right) p_i \tag{14}$$

where the last equality follows from the fact that the elasticity of demand for a particular product is simply the negative of the elasticity of substitution parameter assuming that the overall effect of a price change on the price index is negligible. This gives the familiar set-up where optimal price is a constant mark up over the variable cost.[21]

Product $i$ is imported only if the operating profit exceeds the fixed cost:

setting behaviour would also be more complicated as it has to take into consideration all other cross-price elasticities.

[21] If a change in price were to have an effect on the overall price index, the mark up would depend on both the variable cost and individual product price elasticity. Without information on cost or elasticities, it would not be possible to separate both effects.

$$(p_i - c_i)x_i \geq F_i \tag{15}$$

Assuming products are drawn from a distribution of variable and fixed costs, only products which satisfy the above free entry condition are imported. In equilibrium, quantity demanded has to be equal to quantity supplied. Assuming all consumers are identical this condition is equivalent to:

$$\frac{F_i}{p_i - c} = \frac{\beta_i m}{P}\left(\frac{P}{p_i}\right)^{\sigma} \tag{16}$$

The model highlights the challenges in identifying changes in consumers' preference. An exogenous decline in variable cost or fixed cost gives a similar effect as an increase in the preference parameter, $\beta$. Both effects lead to an increase in quantity demanded and variety of products being consumed.

# B Appendix: Data

Table B1 shows the regions included in each dataset and sample. The baseline OLS regression corresponds to the NHGIS 2010 dataset. The main instrumental variable results uses the NHGIS 1980 foreign-born shares as an instrument. Regression results using NHGIS 1970 and IPUMS 1970 are presented in Appendix C. While Africa and South America are included in the construction of the TF-IDF scores, they are excluded from the regression analysis as relatively few recipes or books from the region were used.[22]

Table B1: Regions Included by Dataset and Sample

| **Dataset** | **Sample** | **Regions** |
|---|---|---|
| NHGIS 2010 | Full | Caribbean, Central America, China, Eastern Europe, France, Germany, Greece, India, Italy, Japan, Korea, Middle East, Other Southeast Asia, Philippines, Scandinavia, Spain, Thailand, Vietnam |
| | Asia | China, India, Japan, Korea, Middle East, Other Southeast Asia, Philippines, Thailand, Vietnam |
| NHGIS 1980 | Full | Caribbean, Central America, China, Eastern Europe, France, Germany, Italy, Japan, Korea, Philippines, Scandinavia, Vietnam |
| | Asia | China, Japan, Korea, Philippines, Vietnam |
| NHGIS 1970 | Full | Caribbean, Central America, Eastern Europe, France, Germany, Italy, Middle East, Northeast Asia, Scandinavia, Southeast Asia |
| | Asia | Middle East, Northeast Asia, Southeast Asia |
| IPUMS 1970 | Full | Caribbean, Central America, China, Eastern Europe, France, Germany, India, Italy, Japan, Korea, Middle East, Philippines, Scandinavia, Vietnam |
| | Asia | China, India, Japan, Korea, Middle East, Philippines, Vietnam |

[22]Nonetheless, the results are robust to including both regions.

Table B2 shows a sample of products by region and score bands as generated by the TF-IDF algorithm. While there might be some misclassification (especially for products with scores lower than 0.5), the list of products seems quite intuitive.

Table B2: Sample of Products by Regions and Score Bands

| Product's score | **Africa** | **Caribbean** | **Central America** | **China** | **Eastern Europe** |
|---|---|---|---|---|---|
| Score <= 1 | mild moroccan fish | navy bean | pinto bean | spicy kung bowl | beef cholent kugel kishka |
| | roasted split turkey breast | baked navy bean | chile tamale | spicy mongolian bowl | string bean |
| | spicy moroccan fish | aromatic | cheesy hashbrown | spicy szechuan ramen | string bean potato |
| Score <= 0.5 | yellow rice | butterflied shrimp | chilies | hoisin sauce | roast duck sauce |
| | regular yellow rice | island getaway seasoning | rotisserie oven roast seasoning | miso hoisin sauce | smoked bacon |
| | lamb stew | shrimp island lime | oven chicken glaze | rakkyo scallion | sour cream chive potato |
| | **France** | **Germany** | **Greece** | **India** | **Italy** |
| Score <= 1 | watermelon rind | buttery rice | rotini tomato basil | kidney bean | gnocchi potato |
| | chestnut puree | assorted | grape leaves | rice pilaf | manicotti |
| | gratin potato | sausage hash canned | beef burger stew | madras lentils | gnocchi |
| Score <= 0.5 | herbs | chervil flakes | macaroni | cayenne | vermouth wine |
| | spice herbs | chervil leaves | elbow macaroni | chili cayenne powder | beef portobello |
| | salad herbs | chervil | shell macaroni | mini whole grain pasta | roast beef gravy |
| | **Japan** | **Korea** | **Middle East** | **Other Southeast Asia** | **Philippines** |
| Score <= 1 | sushi wrap rice bowl | barbecued beef pork | bean rice | peanut satay sauce | beef steak pepper |
| | umeboshi plum | braised beef chili | pinto bean rice | satay sauce | beef steak dinner |
| | sushi ocean crab roll | miso soybean paste | savory bean rice | chile pepper | supreme sushi piece |
| Score <= 0.5 | broiled steak seasoning | rice soup powder | mughlai kofta rice | mild navratan kurma | shrimp spring roll |
| | flame broiled cheese beef | pork napa cabbage dumpling | white beans | mild potato spinach rice | spring roll |
| | flame broiled fajita chicken | seasoned rotisserie | great northern white beans | mild cstnb | spring roll wrap |
| | **Scandinavia** | **South America** | **Spain** | **Thailand** | **Vietnam** |
| Score <= 1 | swedish cream | peruvian bean | spanish style rice | thailand fragrant rice | dragon roll |
| | rock roll berry roll | vino seco wine | pork brains canned | fragrant rice | dragon sauce |
| | rutabaga | white vino seco wine | canary bean | buffalo style | vietnamese noodle |
| Score <= 0.5 | raisin gcmgbl medley | santa style beef | lobster rangoon | tiger sauce | spicy grass chili rice |
| | raisin crispies | mongolian style beef | seafood shrimp lobster newberg sauce fillo | tiger seasoning | grass rice |
| | country style dijon mustard | style beef | lobster cake maine | sticky rice | chix noodle soup |

# C Appendix: Robustness Checks

## CBSA Instead of Counties as the Unit of Analysis

The use of county level data as the unit of analysis is based on the assumption that potential peer effects on consumption occur at the county level. If such spillover effects happen at a smaller geographical unit of analysis, the analysis at the county level could be interpreted as averaging over the smaller areas and the standard errors might be larger than expected. On the other hand, if such spillover effects occur at a broader unit of analysis, the results presented could potentially overstate the level of significance. Table C1 repeats the OLS and IV regression using core based statistical areas (CBSA) instead of counties as the unit of analysis. CBSA is a geographical grouping of counties that contains at least 10,000 people anchored by an urban center. The estimated effects are similar to the results obtained using county level data.

## Alternative Datasets and Region Groupings

Table C2 presents two additional set of IV regressions using different datasets to construct the instruments. NHGIS 1980 is the original dataset used in the paper. IPUMS 1970 contains a more detailed breakdown of foreign-born shares by region but at the expense of much fewer counties. NHGIS 1970 has data available for most regions but data on foreign-born shares is aggregated over broad geographical regions. Once again, the coefficient on foreign-born share is positive and significant for the full sample but negative or not significant when restricting to the sample of only Asian regions.

Table C1: Regression Estimates at the CBSA Level

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | | | Exp Share | |
| | OLS | OLS | IV | IV |
| | (1) | (2) | (3) | (4) |
| 2010 FB Share | 0.2301*** | −0.0598 | | |
| | (0.0787) | (0.0430) | | |
| Fitted 2010 FB Share | | | 0.3096*** | −0.1567* |
| | | | (0.0606) | (0.0840) |
| First Stage F-stat | | | 341.51 | 117.37 |
| Sample | Full | Asia | Full | Asia |
| Region FE | X | X | X | X |
| CBSA FE | X | X | X | X |
| Observations | 646,816 | 363,834 | 483,744 | 201,560 |

Notes: Two-way standard errors clustered by core-based statistical areas (CBSA) and region in parentheses. OLS sample consists of 18 regions, OLS Asia sample consists of 9 regions, IV full sample consists of 12 regions, IV Asia sample consists of 5 regions. *p<0.1; **p<0.05; ***p<0.01.

Table C2: Robustness Tests (Different Datasets)

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Exp Share | | | | | |
| | IV | IV | IV | IV | IV | IV |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Fitted 2010 FB Share | 0.2762*** | −0.1337** | 0.2562* | −0.0569 | 0.4641*** | 0.0855 |
| | (0.0480) | (0.0669) | (0.1362) | (0.0516) | (0.1324) | (0.1421) |
| First Stage F-stat | 105.99 | 58.71 | 11.14 | 34.88 | 4.94 | 65.46 |
| Sample | Full | Asia | Full | Asia | Full | Asia |
| Dataset | NHGIS 1980 | NHGIS 1980 | IPUMS 1970 | IPUMS 1970 | NHGIS 1970 | NHGIS 1970 |
| Region FE | X | X | X | X | X | X |
| County FE | X | X | X | X | X | X |
| Observations | 595,920 | 248,300 | 202,244 | 101,122 | 496,550 | 148,965 |

Notes: Two-way standard errors clustered by county and region in parentheses. Baseline NHGIS 1980 IV full sample consists of 12 regions, IV Asia sample consists of 5 regions. IPUMS 1970 IV full sample consists of 14 regions, IV Asia sample consists of 7 regions. NHGIS 1970 IV full sample consists of 10 regions, IV Asia sample consists of 3 regions. The results of specification three and four which use the NHGIS 1970 dataset are also calculated based on a revised set of TF-IDF scores generated using the broader region groupings. *p<0.1; **p<0.05; ***p<0.01.

# References


Altonji, J. G. and D. Card (1991). The effects of immigration on the labor market outcomes of less-skilled natives. In *Immigration, trade, and the labor market*, pp. 201–234. University of Chicago Press.

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics 30*, 98–108.

Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance 59*(3), 1259–1294.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131*(4), 1593–1636.

Bellini, E., G. I. Ottaviano, D. Pinelli, and G. Prarolo (2013). Cultural diversity and economic performance: evidence from european regions. In *Geography, institutions and regional economic performance*, pp. 121–141. Springer.

Birch, L. L. (1999). Development of food preferences. *Annual review of nutrition 19*(1), 41–62.

Borjas, G. J. (2003). The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market. *The quarterly journal of economics 118*(4), 1335–1374.

Brock, W. A. and S. N. Durlauf (2001). Interactions-based models. *Handbook of econometrics 5*, 3297–3380.

Bronnenberg, B. J., J.-P. H. Dubé, and M. Gentzkow (2012). The evolution of brand preferences: Evidence from consumer migration. *The American Economic Review 102*(6), 2472–2508.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *ILR Review 43*(2), 245–257.

Card, D. (2001). Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics 19*(1), 22–64.

Dixit, A. K. and J. E. Stiglitz (1977). Monopolistic competition and optimum product diversity. *The American Economic Review 67*(3), 297–308.

Einav, L., E. Leibtag, A. Nevo, et al. (2008). *On the accuracy of Nielsen Homescan data*.

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

Gould, D. M. (1994). Immigrant links to the home country: empirical implications for us bilateral trade flows. *The Review of Economics and Statistics*, 302–316.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Hunt, J. and M. Gauthier-Loiselle (2010). How much does immigration boost innovation? *American Economic Journal: Macroeconomics 2*(2), 31–56.

Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28*, 11–21.

Kerr, S. P. and W. R. Kerr (2011, January). Economic Impacts of Immigration: A Survey. NBER Working Papers 16736, National Bureau of Economic Research, Inc.

Kerr, W. R. and W. F. Lincoln (2010). The supply side of innovation: H-1b visa reforms and us ethnic invention. *Journal of Labor Economics 28*(3), 473–508.

Kibriya, A. M., E. Frank, B. Pfahringer, and G. Holmes (2004). Multinomial naive bayes for text categorization revisited. In *Australian Conference on Artificial Intelligence*, Volume 3339, pp. 488–499. Springer.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies 60*(3), 531–542.

Manson, S., J. Schroeder, D. Van Riper, and S. Ruggles (2017). IPUMS National Historical Geographic Information System: Version 12.0 [Database]. Minneapolis: University of Minnesota.

Nestle, M., R. Wing, L. Birch, L. DiSogra, A. Drewnowski, S. Middleton, M. Sigman-Grant, J. Sobal, M. Winston, and C. Economos (1998). Behavioral and social influences on food choice. *Nutrition reviews 56*(5), 50–64.

Ottaviano, G. I. and G. Peri (2006). The economic value of cultural diversity: evidence from us cities. *Journal of Economic geography 6*(1), 9–44.

Rauch, J. E. and V. Trindade (2002). Ethnic chinese networks in international trade. *Review of Economics and Statistics 84*(1), 116–130.

Rozin, P. and T. A. Vollmecke (1986). Food likes and dislikes. *Annual review of nutrition 6*(1), 433–456.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek (2015). Integrated Public Use Microdata Series: Version 6.0 [dataset]. Minneapolis: University of Minnesota.