An ethnic classification using machine learning and surname data and its applications to social science

by

Joomi JUN

Dissertation

submitted to the Department of Informatics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

The Graduate University for Advanced Studies, SOKENDAI
March 2023

Abstract

People leave their birthplace and settle in another land for political, religious, or economic reasons. Thus, people from different origins and political and cultural backgrounds mix, making society more diverse. To understand society, one needs to first understand the difference and characteristics of members of society. In this study, I focus on ethnicity to understand social phenomena. Ethnicity combines physical, cultural, linguistic, and other characteristics. Its definition and level of subdivision differed by country. Therefore, quantitative analysis was difficult. To address this limitation, this study uses surnames to identify individual's ethnicity and applies research on social phenomena.

I develop a classifier that could verify the origin of people using business people's surname from ORBIS dataset and Recurrent Neural Networks. Surname-origin classifier identifies individuals' origin from his/her surname. This classifier can identify 77 origins. A test using Olympic data yielded an average accuracy of 66.8%. In addition, compared to the human judgment result obtained through MTurk, origin was classified similar level of local people.

Using surname-origin classifier, I analyze the ethnic diversity in sociology and the ethnic homophily in economics. For investigating the ethnic diversity in sociology, I first analyzed the ethnic diversity of each country with calculating the entropy of the origin composition. Through this, it is possible to compare ethnic diversity between countries with an overwhelmingly high proportion of a dominant ethnic group, such as Japan, and countries like Canada, which do not.

Next, I focus on the social groups in the United States and showed how the racial composition differed by group. Based on the racial proportions of US demographics, it is confirmed that Asians occupy a relatively high proportion in the economic and physician groups, and non-Hispanic whites occupy a higher proportion in public figures than statistics.

Finally, I analyze the spatial distribution of origin in African continent and Europe countries. The racial composition of all subjects of analysis is complex, and they are in an environment where it is difficult to analyze them statistically. In this study, origins are

classified by surname, and regions with similar origin distributions are grouped by Jensen-Shannon Divergence. In the case of the African continent, it was divided into four groups. This yielded similar results to subdividing the African continent by language and religion.
For analyzing the ethnic homophily in economics, I focus on the international trade. First, I analyzed the frequency of transactions between ethnic groups focusing on companies in the United States. Considering the majority ethnic group of each company and using the conditional probabilities to visualize the frequency of transactions between ethnic groups, I found strong ethnic homophily, especially in the Asian and Middle East ethnic groups.
Next, using the gravity model to measure the statistical significance of ethnic factors in international trade, I analyzed the global trade using the WTFC dataset, to find a positive ethnic factor effect, although this differed by country. Although there was no significant ethnic factor effect in the English (GB) and European (DE) groups, positive effects could be observed in the Chinese (CN) and Korean (KR) groups. I also found significant ethnic factor effects even after removing the language barriers. A preference could be found between identical ethnic groups not in domestic, but in international trade.

Finally, I introduce case study which interethnic trade is activated when an obstacle occurs in international trade. For this, I used the US trading data obtained from FactSet and the PIERS bill of lading data obtained from IHS, and analyzed how ethnic networks in international trade are activated around specific events such as the U.S.–China trade war and the Arab Spring. This study assumed that ethnic linkage can be used to overcome the existing difficulties in international trade. I could thus conclude that exports from China to the United States decreased slightly compared to the average during the 2018 US–China trade war. However, when a Chinese executive was employed in the company of the consignee during the same period, the company's imports from China increased compared to the average. Thus, ethnic networks can be activated to overcome inter-country trade disturbances. During the Arab Spring period, Arabian's interethnic trade was also observed to be active.

This study is meaningful in that it examines the social science problems of ethnic composition using massive data and machine learning techniques adopting the informatics approach. The study is crucial in that it addresses the ethnic diversity and homophily issues that affect social phenomena.

This study can be extended to other fields, for example, to resolve ethnic conflicts and health care issues by ethnic characteristics. It can also be applied to cryptocurrency market, which is expected to be active in interethnic transactions because language dependence becomes stronger to obtain information. I leave these tasks to future works.

Acknowledgement

I want to express my gratitude to the many people who helped and inspired me to complete this dissertation. First, I would like to thank my supervisor, Prof. Mizuno, for instructing me in detail on the path of research and teaching me how to think anew during a lengthy doctoral course. Also, thank you for encouraging and guiding me when I got stuck in research and when I felt difficulties in the process of finishing. Prof. Takeda awakened me to the point of view I hadn't thought of during research and informed me of my deficiencies. Prof. Miyao gave me the necessary parts for a complete research thesis. Prof. Ohnishi presented various perspectives and ideas in research presentations such as seminars. Prof. Okada, for letting me know in detail about a field I was unfamiliar with and for allowing me to broaden my perspective. Thank you so much.

Thank you to the laboratory members, Mr.Saito, Ms.Jin, Mr.Horikomi, Mr.Odaka, and Dr.Doi, who provided ideas necessary for research and positive stimulation through the seminar. Thanks to this, I was able to widen my field of view, which could have been narrow. Thank you, Ms.Minami and Ms.Takedomi, who kindly supported me in every detail, such as research and graduate school life.

I would also like to express my sincere gratitude to SOKENDAI for helping me concentrate on my research with administrative and financial support.

And I am infinitely grateful to my family, Jaechul JUN, Gunsook LEE, Kiyou JUN, Bok, and Taekuek, for the love and support they have always sent me.

Joomi JUN
Jan. 2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

People cross borders for various reasons. They leave their birthplace and settle in another land for political, religious, or economic reasons. Thus, people from different origins and political and cultural backgrounds mix, making society more diverse. To understand society, one needs to first understand the different members of the society. This study analyzes the social phenomenon based on member ethnicities.

First, we need to define ethnicity as a social factor to understand the social phenomena. Therefore, I will briefly define ethnicity and similar terms used in this study.

Ethnicity: Classification by complex factors such as biology, geography, culture, religion, and linguistic aspects.
Nationality: Legal definition of belonging to a particular political group of nations.
Origin: Birthplace of a family root or the primary nationality of a specific surname.
Race: Classification by physical characteristics such as skin color, hair color, and appearance. This often transpires as biological classification.

For a man named Mori Shinichi with American nationality, one can assume that his nationality is American, his origin is Japanese, his race is Asian, and his ethnicity is Japanese.

This study focuses on ethnicity while analyzing the background of social members. Ethnicity is closely linked to several aspects such as language, culture, and religion. To track the origin of ethnicities in society, this study uses the surname data of people, especially those in the business field. The study is based on the literature on sociology [JJ20], [JJ21b] and economics, [JJ21a] and uses an ethnic classification model with surname data.

With the increased use of computer technology, studies classify races, ethnicities, and nationalities by their name data using machine learning and statistical models. Name data contain a variety of information about individuals. For example, the first names of individuals often provide information about gender, historical trends, cultural backgrounds, and nationality. Surnames provide information on the roots of family systems and the origin of their ethnicity. Name data, with a unique structure based on ethnicity and historical and national backgrounds, enable one to obtain and analyze various social data that cannot be found through statistical investigation.

Using the recurrent neural network (RNN) method, I build a surname origin classifier using the recorded data of surnames and nationalities of business persons to predict their origin. As related studies do not specifically define the categories of ethnicity in detail, I cluster the surname vectors of each origin with similar ethnic structure. Thus, I improve the accuracy of classification, especially with similar ethnic compositions affected by language or cultural environment.

Using the surname origin classifier, I then analyze the following sociological applications. First, I show that diversity of origins in each country can be analyzed with surname data alone. Next, I analyze how the races in each social group (physicians, public identities, business people) in the United States, a representative multi-ethnic nation, differ from the US demographics. I also visualize the spatial distribution of origins of the African continent and European countries.

I conduct a detailed analysis of ethnic linkage in the economic field. I determine the role of ethnic linkage in international trade by examining whether international economic transactions activate the ethnic network in each country. Ethnic networks play an important role in international trade, which involve diverse entities. Ethnic networks can effectively reduce the cost of information barriers and unreliable contracts in international trade. Related studies also have recognized the importance of people involved in international trade and have studied their influence. However, as there is a limitation in terms of classifying ethnic groups and obtaining such data, this study conducted a detailed research around only one ethnic group. The study considers the ethnic linkage and influence between countries and groups using ethnic classification techniques with surname data, and examines whether this relation is statistically significant.

This study is meaningful in that it examines the social science problems of ethnic composition using massive data and machine learning techniques adopting the informatics approach. It constructs a simple prediction model that can solve racial division problems and estimate the impact of ethnic linkage on economic networks. The study is crucial in that it addresses the ethnic diversity and homophily issues that affect social phenomena. Adopting this approach, I present a case in sociology where this method can be applied. I also quantitatively compare and analyze the ethnic dependence of all ethnic groups—not just one—using surname data. I determine the ethnic groups that are heavily influenced by ethnicity in international trade and then, using the gravity model, I measure the statistical significance of ethnicity factors in international trade.

The remainder of this paper is organized as follows. Related studies will be introduced in Section 2. Section 3 introduces the dataset, while Section 4 explains how to build the surname origin classifier and presents its test results. Sections 5 and 6 provide examples of an application study in sociology and economics, respectively. Section 7 presents various opinions on ethical issues related to ethnic studies. Finally, Section 8 concludes the paper.

Chapter 2

Related work

Ethnic and racial classification studies using name data have been conducted in various fields. First of all, the most active area is the health and biomedical field. They actively research the medical characteristics expressed in specific ethnic groups using matching data of surnames and geocoded information.

In [BRS10], they use surname lists for identifying cohorts of ethnic minority patients, predominantly South Asian and Chinese origin. [DSL97] estimated national age-and sex-specific nontraumatic hip fracture incidence rates for Chinese Americans, Japanese Americans, and Korean Americans using distinctive names. [MNE08] and [SFD13] estimate race/ethnicity using combined surname and geocoded information with the Bayesian approach. In [SLS99], they try to classify Hispanic ethnicity using surname lists for the statistics of ethnic-specific cancer. [HQ06] develops and validates surname lists of Chinese ethnicity using a survey. [MG02] uses South Asian surnames and first names for researching Coronary artery disease, which affects a significantly larger proportion of Canadians of South Asian origins than other origins. [IIW06] uses the Census Spanish surname list to improve the identification of Hispanic women in medicating administrative data. [Gor99] examines the influence of individual-level characteristics on low birthweight risk for each race using a linked birth and death data set. [FL11], [GJM07] established name-based ethnicity classification methodologies 'Onomap' and 'Nam Pehchan' for addressing health inequalities between ethnic minorities and the general population, respectively. [AMES10] validates an empirically based probabilistic Arab name algorithm for identifying Arab-Americans. [ECW10] also uses a name list for identifying Asian subgroups in the clinical data source. In [KLS04], the authors create an Arab/Chaldean surname list for studying the cancer distribution in the Arab ethnic groups. [DS10] focuses the ethnicity on examining the relationship between ethnicity and tolerance of hypertension medications. Ethnicity has been treated as a significant factor in public health care, and in 2019, a more improved surname-geocoded classifier was released according to racial classification in the United States [AH19].

In social science, the research tries to understand ethnic effects and social phenomena using surnames in various fields such as politics, education, and the economy. [IK16] uses a surname list with geocoded information to predict individual-level ethnicity from voters' registration records. [Col05], [Sad20] use similar data, and they study the bloc voting and polarization in each race. In [DSL02], the authors determined ethnicity by combining race, place of birth, surname, and given name from Social security Administration files to analyze death probabilities for the Asian subgroups. [Ahm10] shows their experiment results that investigate the possible incidences of discrimination using surnames. In [DD11], they use a Korean surname list to determine Korean immigrants and measure immigration stress. [AV01] uses grandparents' surnames from primary schools, and they analyze the relationship between two ethnic minorities. In [JB19], they analyze the ethnic homophily of mover flows in Glasgow. [Cry15] identifies the Irish in 18th-century London by surname analysis. [YS19] investigates surname affinities among areas of China by constructing a spatial network and community detection. In [Daf19], the author examines the ethnic segregation of immigration groups in the U.S. based on the linguistic origin of the surname.

Statistics and Computer science focus on the classification problem using surname data rather than applied research. [AM11], [LKK$^+$17], [Mat14], [AA09] build a statistical model to estimate the origin/ethnicity with the labeled training datasets. study [AM18] develops a surname origin classifier based on a data-driven typology. Census data, phone book data, Olympic data, and Wikipedia data used in these related studies are still widely used data for ethnic classification using surnames. Various methodologies have also been studied. [AA09] applied Hidden Markov Models(HMM) and decision trees, [VIT16] applied logistic regressions, by the Bayesian Approach [JC10], a machine learning method [MP11], and [AM18] applied the Naive Bayes model. [LKK$^+$17] built a classifier using Recurrent Neural Networks(RNNs).

As such, ethnic studies using surnames are being conducted in various research fields. However, due to the data that can be collected, it is impossible to rely on surname lists and deal with the problem of classifying a specific ethnic group. In addition, active research has not yet been sufficient in social science, which is the primary focus of this paper. This paper presents a method for classifying ethnic groups subdivided with a large amount of surname data, which has not been done in related studies and also shows the results of social scientific applied research.

Chapter 3

Dataset

In this chapter, I will explain the datasets I used for this study. The data is divided mainly into 2 types of data. First is the name data and the second is the trading data. I will explain what datasets are included in each data type and briefly explain the purpose for which the dataset will be used.

3.1 Name data

A name is a person's most basic index of information. Names are created and used by characteristic systems of society and culture, and so they can suggest gender, region, and cultural background. From first names, we can often infer a particular region, gender, trends of age, etc. From family surnames, we can recognize roots that resemble the origin of ethnicity. For example, "Van Gogh" is a Dutch-specific family name that means "from/of Gogh," and the roots of the painter's family are clearly located in the Gogh region. Name data, which have a unique structure with ethnicity, historical, and national backgrounds, enable us to obtain and analyze various social information that cannot be found in statistical investigations. Therefore, since the accumulation of name data, it has been paying attention in various study fields like: Anthropology([San00] [JCS02] [IB96]), Biology and Health([DSLG96] [AN86] [MHDLD13]), Language linguistics([BB13] [Ker11]), Sociology([EL06] [BM09] [MT09]), etc.

I wanted a reliable name dataset to study the influence of ethnicity based on it. Name data such as phone books may also be used. But phone book is limited to regions and countries; therefore, it is not suitable for analysis and utilization in a wide range. The ORBIS dataset, FactSet dataset, IHS dataset were very reasonable for my research purpose.

The information about global companies and people related to the company was reliably built. Details of the two datasets and additional datasets used to analyze ethnic influence are discussed in this chapter. In this chapter, I describe the datasets commonly used in the entire manuscript. The data used auxiliary in each Section will be explained separately.

3.1.1 ORBIS dataset

In this study, large-scale reliable name data is required for learning, and the reason adopted is the ORBIS dataset. ORBIS is a database provided by Bureau Van Dijk, which operates to provide database solutions that can support business decisions. The database can be obtained by requesting the Bureau Van Dijk Corporation [ORB]. ORBIS provides data for listed and unlisted companies around the world. The information of the company included is as follows [ORB].

<Information about company>

- financial strength metrics and projected financials
- scores on companies with limited financials
- associated news and independent research
- extensive corporate ownership structures and beneficial ownership information
- original as-filed documents and document ordering services
- data on individuals associated with companies
- global M&A deals and rumours
- marine vessels data
- ESG reputational risk ratings and metrics
- public tenders data
- PIEs (Public Interest Entities)
- patents and intellectual property
- royalty agreements
- PEPs and sanctions

The database provides information for about 400 million companies (99% are unlisted companies). This includes incorporation date, address, mail address, website URL, and phone number. Each company is searched by a unique ID called BvD given by Bureau Van

Dijk. In addition, Local ID numbers and LEI (Legal Entity Identifiers) are also provided to search the company. The database also includes investment information (shareholding, subsidiary information, direct & indirect ownership, ultimate owners, independence indicators, corporate groups, company tree diagrams, beneficial owners, etc.).

ORBIS database also provides the data of the relevant people within or connected to companies: like as advisors, auditors, board members, directors, senior managers, other contacts. In other words, management positions related to the primary decision-making power of the corporate business are mainly provided. This data includes the information of a person as follows.

<Information about person>

- Company name
- BvD ID
- Full name
- UCI(Unique ID of the person)
- Original Job Title
- Arrival date
- Retirement date
- current / past
- Salary (USD)
- Total remuneration (USD)
- Individual/Company
- Gender
- Date of Birth
- Age
- Nationality

Examples of some data of the above items used in this study are shown in the 3.1. Some contents were covered to protect personal information.

Table 3.1: Sample data abstracted from ORBIS dataset.

Title	Full name	Birthplace (city)	Nationality	Birth year
Mr.	Alka*** A	RIYADH	Saudi Arabia	19*6
Mr.	Dar*** Franceschetto	BARBARANO VICENTINO	Italy	19*2
Ms.	Mar*** Ivascu	NA	Romania	19*9
Ms.	Kar*** Raloff	NA	Germany	19*9
Mr.	Khu*** yan	NA	China	19*9

The data include data in each country's primary language, but this study only deals with data written in the alphabet.

I use ORBIS 2016 dataset. This dataset include 37 million names and nationalities of company executives and individual shareholders in 203 countries.
Registered data from 203 countries have very high deviations. Russian Federation has the largest data quantity (total of 6,957,186 people). Kiribati has the least, which contains only two people. For machine learning and analysis, 77 countries with data of more than 3,000 people are selected. Figure 3.1 shows 77 selected countries and the number of people(company executives) registered in those countries. Since surname data is important in this study, 77 countries were selected based on the person. However, there is a slightly different aspect in the number of companies. Of the 77 countries, the country with the largest number of companies registered in the United States (20,973), and Algeria (5) with the least number of companies registered (Figure 3.2). This seems to differ depending on how many are registered as private businesses. In the ORBIS 2016 dataset, companies in each country except Japan have up to 60 executives reported. In Japan, about 100 executives are registered, which differs depending on the data survey method. Although not used in this paper, ORBIS data for 2021 were also confirmed. The maximum number of registered executives has increased significantly in the data, and there are about 3,000 companies (public institutions).

ORBIS dataset is used for creating the classification model for Chapter 4 and application analysis using the names and ethnicities of Chapter 5.

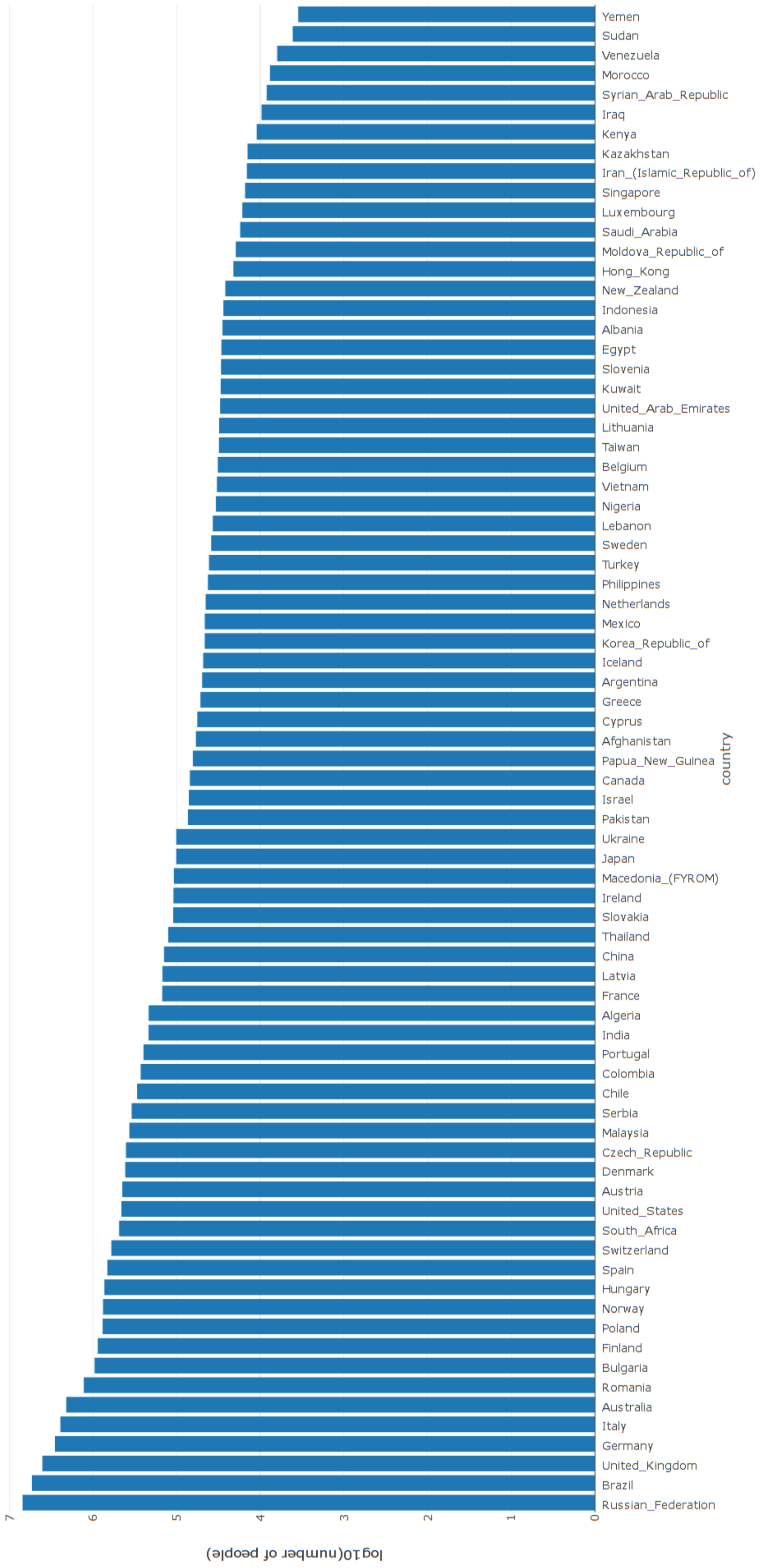


Figure 3.1: 77 Countries and number of people (company executives) in ORBIS 2016 dataset.

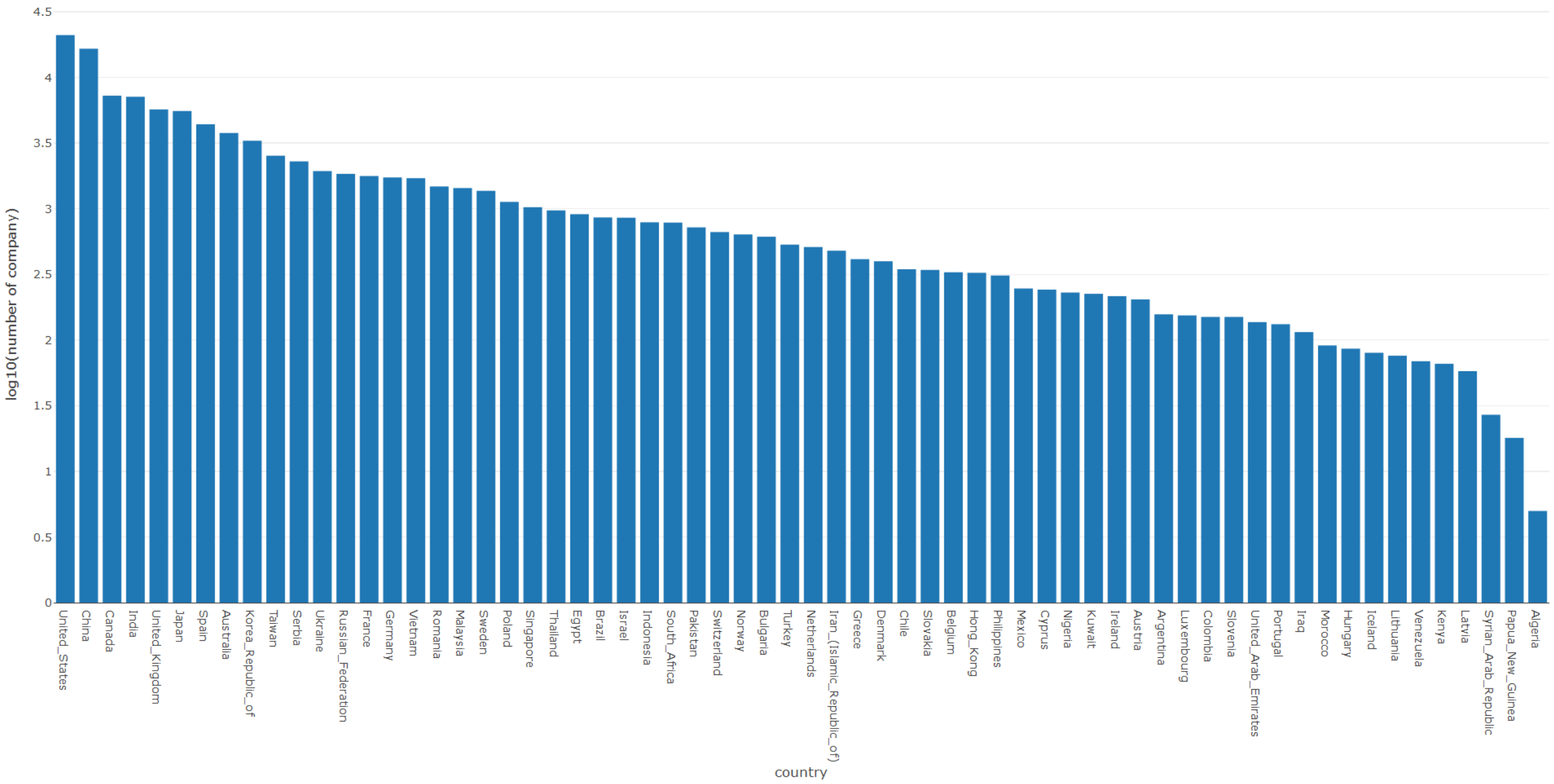


Figure 3.2: 77 Countries and number of Companies in ORBIS 2016 dataset.

3.1.2 DowJones watch list dataset

Dow Jone's Watchlist identifies high-risk third parties to assist in complying with global Anti-Money Laundering regulations. This includes enhanced and consolidated data on people and entities regarding regulatory sanctions and other official lists, Politically Exposed Persons (PEPs), Relatives and Close Associates (RCAs), and Special Interest Persons (SIPs) who have been involved in a legal process in relation to defined criminal categories, and Sanctioned by Control and Ownership entities. The Watchlist dataset includes items such as Sanctions and other lists, Politically exposed persons, Special interest and reputationally exposed persons, and Sanctions control and ownership. The dataset contains information on 2 million people in 219 countries. The number of identities for each country is shown in the figure 3.3. I use this dataset to analyze the proportion of public figures in the United States by race in the application analysis in Chapter 5.

3.1.3 Olympic participants dataset

This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. The author of this dataset scraped this data from www.sports-reference.com in May 2018. This dataset contains 134732 names and 230 nationalities. Table 3.2 shows example of the dataset. The data also includes information about the athlete's height, weight, and medals, which are omitted from the example. This dataset is published in the Kaggle[Gri18]. I use this dataset to evaluate the classifier's performance in Chapter 4.

Table 3.2: Example of Olympic data

id	name	nationality	sex	dob	sport
736041664	A Jesus Garcia	ESP	male	10/17/69	athletics
532037425	A Lam Shin	KOR	female	9/23/86	fencing
435962603	Aaron Brown	CAN	male	5/27/92	athletics
521041435	Aaron Cook	MDA	male	1/2/91	taekwondo
33922579	Aaron Gate	NZL	male	11/26/90	cycling
173071782	Aaron Royle	AUS	male	1/26/90	triathlon

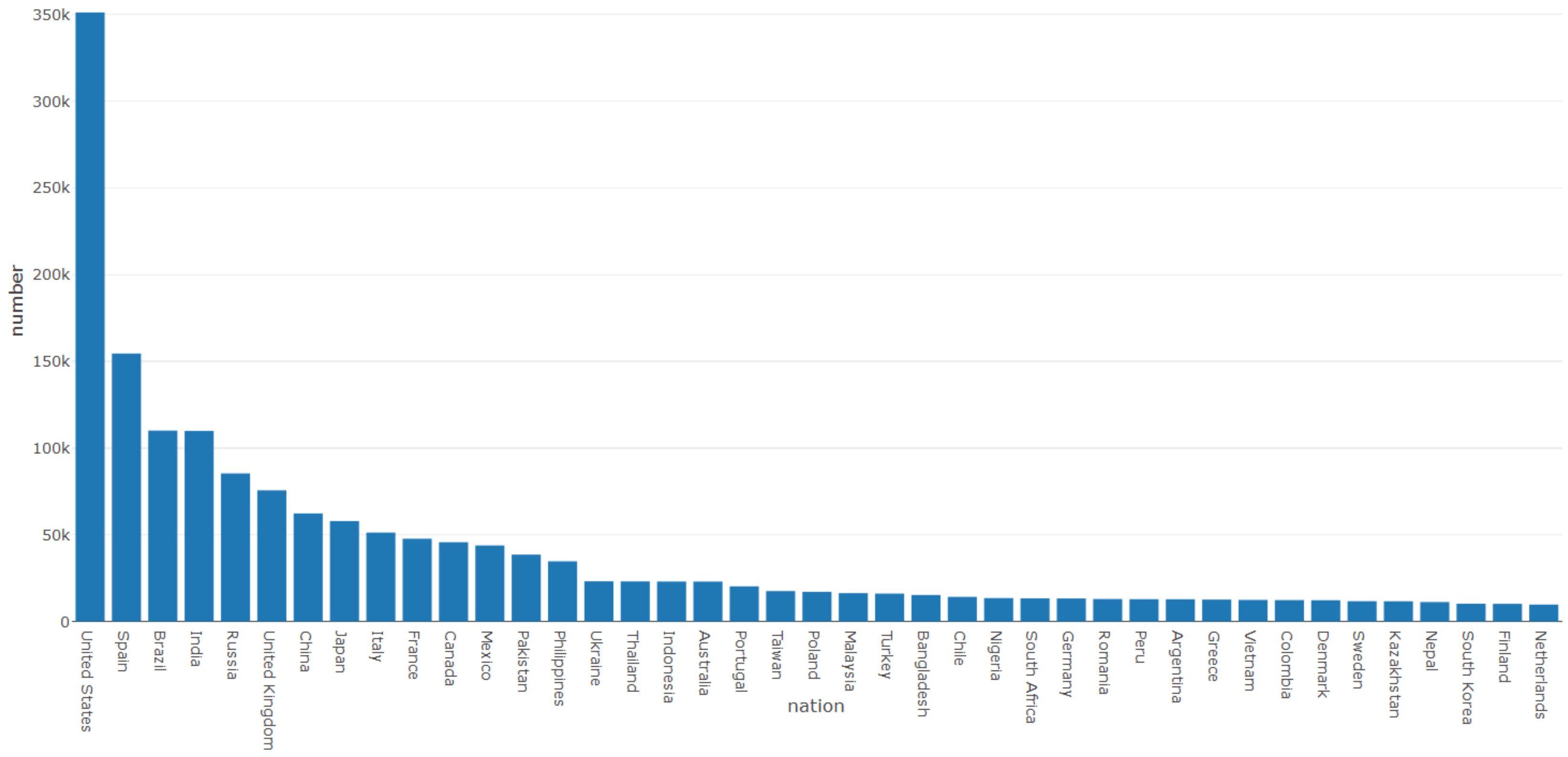


Figure 3.3: The number of registered persons for each nation in the Dow Jones Watchlist dataset indicates the case of 41 countries with an average of more than 9731.8 people out of 219 nations.

3.1.4 U.S. Census dataset

The U.S. Census Bureau publishes a list of the surnames most used by Americans and the frequency for each surname every ten years. I use this dataset for analyzing the racial composition of US social groups in Chapter 5. As described in detail in the 2000 data, 151671 unique names were extracted from over 6 million people. For example, the surname Smith, which appears most frequently, was the surname of 2.3 million people, representing 9% of the total population. They also list six racial groups (White, Hispanic, American Indian and Alaska Native, Black, Asian Pacific Islanders, and Two or more races). They include information on the most common surnames for each race. Data can be downloaded from [USC]. Table 3.3 shows example of the dataset.

Table 3.3: Example of 2010 US Census data

name	rank	count	prop100k	cum	w	b	a	an	2r	h
SMITH	1	2442977	828.19	828.19	70.9	23.11	0.5	0.89	2.19	2.4
JOHNSON	2	1932812	655.24	1483.42	58.97	34.63	0.54	0.94	2.56	2.36
WILLIAMS	3	1625252	550.97	2034.39	45.75	47.68	0.46	0.82	2.81	2.49
BROWN	4	1437026	487.16	2521.56	57.95	35.6	0.51	0.87	2.55	2.52

※column head of Table 3.3 and definition

name : Surname

rank : National rank in 2010

count : Frequency: number of occurrences nationally in 2010

prop100k : Proportion per 100,000 population for name

cum : Cumulative proportion per 100,000 population

w : Percent Non-Hispanic White Alone

b : Percent Non-Hispanic Black or African American Alone

a : Percent Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone

an : Percent Non-Hispanic American Indian and Alaska Native Alone

2r : Percent Non-Hispanic Two or More Races

h : Percent Hispanic or Latino origin

3.1.5 American physicians dataset

This data is from the "Codebook for Replication Materials for “The Political Polarization of Physicians in the United States: An Analysis of Campaign Contributions to Federal Elections, 1991 Through 2012.” This codebook provides a reference for the database on the political contributions of physicians used in Bonica, Rothman, and Rosenthal (2014). The database combines the National Provider Identifier (NPI) and Unique Physician dentifier Number (UPIN) directories to identify physicians and links gainst Database on Ideology, Money in Politics and Elections (DIME) to identify their political contributions. They provide 4 types of list of files as below.
1. The main database with the merged NPI/UPIN and DIME data. 2. A table that reports the Democratic two-party presidential vote share by congressional district and election cycle. 3. A table listing the total number of individual donors in DIME active in federal elections for each cycle. 4. R data file that stores the trend for the mean cfscore for the entire population of individual donors across election cycles.

This dataset includes total of 1944930 identities. I use this dataset to analyze social group racial distribution in Chapter 5. Table 3.4 shows the example of the data. This dataset contains data such as the physician's full name, address, and donation amount by election period, but in the example, only the ID and surname, gender, and specialty are displayed. This file can be downloaded from [Bon]

Table 3.4: Example of American physicians data

ID	surname	gender	specialty
1679576722	WIEBE	M	ORTHOPEDIC SURGERY
1588667638	PILCHER	M	CARDIOLOGY
1306849450	SMITSON	M	DIAGNOSTIC RADIOLOGY
1669475711	SUNG	F	PEDIATRICS
1659374601	OBERDICK	F	FAMILY MEDICINE

3.2 Trading data

3.2.1 FactSet shipping dataset

FactSet is a company that provides global information to investment experts. They have more than 40 years of operational experience and more than 37 offices are located around the world to provide services. Similar to the data in 3.1.1., database access can be obtained through the contact of the company. [Facb] An overview of the data provided in FactSet is as follows.

<FactSet database>

- 20 million shipments/transactions
- 700,000 shipper/consignee relationships for 200,000 unique entities in 100,000 entity trees
- 200 US ports and nearly 1,000 non-US ports
- 9,000 vessels

I want to use shipping data and supply chain data as data to check the status of global trade. Data from 2013 to 2018 are used to analyze. Shipping data is a manually written data set based on the bill reported to the Customs Service in U.S. ports. The following items are recorded in this data set:

<Shipping Record> [Faca]

- Transaction ID : unique identifier of each shipping transaction record
- Start date of a record (UTC)
- End date of a record (UTC)
- The FactSet entity ID for the shipper
- The FactSet entity ID for the consingnee(recipient)
- The FactSet entity ID for the carrier(vessel)
- Date on which the record was districuted by the originainating source
- Status of record as of the last time the record was received from the originating source
- Estimated arrival date when the shipment reached its destination
- Actual arrival date

- Entity ID for the originating port for this transaction
- Entity ID for the destination port for this transaction
- The date FactSet content collection first processed the record from the source

<Item Records> : individual records for identified content item for the associated transaction.

- Item ID : unique identifier for each content item record.
- Start date of the record in the feed.
- End date of the record in the feed.
- Container identifier number of the item.
- Total number of items in the entire transaction.
- Full textual description of the content of items provided by the shipper.
- Type of HTS code for the content item.
- The FactSet normalized HTS code.
- Dollar value for the item.
- Weight of the shipment item.
- Total number of pieces of the item in the entire transaction.

The total shipping record is 46,306,411, and the record of the final destination of the U.S. port is about 89% of the total. The rest are trade records for other countries via U.S. ports. The FactSet database also provides information on each company and information on corporate executives separately. A total of 228,598 executive information is included and consists of the personal entity ID, job title, company ID.

The company's information includes SIC codes. In particular, the SIC code I use to analyze the company's sector and remove unnecessary records. Table 3.5 shows how many companies are included by sector. As can be seen from the data, due to the nature of trade transactions, many financial companies, accounting companies, and insurance companies that pay for the transaction of goods have been recorded to handle transaction items primarily. There is also information about trade goods. In this study, data such as HTS code, weight, and price of goods are compared with trade data from the U.S. Census, and how much FactSet data can reproduce U.S. trade data is reviewed.

Table 3.5: Number of companies by sector registered in FactSet

Sector	number of company
Finance Insurance Real Estate	1,459,617
Service	669,393
Manufacturing	548,848
Transportation, Communications, Electric etc.	357,138
Wholesale Trade	164,065
Retail Trade	120,498
Construction	88,428
Mining	50,423
Nonclassifiable	44,570
Public Administration	42,700

Since this dataset also includes information on people of managerial positions affiliated with import or export companies, the country of origin of their surnames can be estimated using Surname-Origin Classifier. I explicitly investigate ethnic dependence in the choice of international trading partners with the transaction data. I removed self-loop transactions and unclear ownership of importing and exporting to clarify the relationship between exporter and importer. So I excluded (1) transactions between parent-subsidiary relationships, (2) transactions in which exporting company's nationality and exporting port's locations are not matching, and (3) transactions recorded as imports by financial companies making importing payments. And I selected companies that have more than three managers for estimating the majority's ethnicity. After the data cleaning, a total of 15,503,150 transactions is used. It's 27.7% of whole transaction data.

3.2.2 FactSet supply chain dataset

This dataset is FactSet's trade relationship information of 31,600 global companies. This information is collected from company annual reports and public news articles worldwide. Companies and people in managerial positions are linked in the dataset. The managerial positions include low-level and top-level managers—for example, a team manager is a lower-level manager, and a president or CEO is a top-level manager. I will call the employees in managerial positions a 'manager' onward. I selected companies that have more than three managers and removed parent-subsidiary relationships. The cleaned dataset includes 479,448 pairs of supplier-customer relationships from 2014 to 2020. I use this data set for analyzing the interethnic trade in Chapter 6. Table 3.6 shows examples of the dataset.

Table 3.6: Examples of the FactSet Supply chain data in 2020

F.ID	Company	nation	trade	F.ID	company	nation
05***	Se*** Ltd.	SG	->	07***	PT ** Energy	ID
05***	Bang *** A/S	DK	->	05***	KO*** Corp.	KR
06***	Holl** Technologies Ltd.	CN	->	0B***	Dal** Industries Ltd.	IN
05***	Ima*** Group Ltd.	GB	->	0F***	ELV*** CJSC	RU

3.2.3 WTFC trade dataset

The World Trade Flows Characterization (WTFC) dataset (provided by CEPII [CEP]) contains trade value (US dollars) of products, HS code of the products, trade type in more than 230 countries. They also provide each country's GDPs and the distance between cities with the highest population distribution of each country. I use the 2016 aggregated trade values by country, GDP, and distance for measuring the significance of the ethnic factor in international trade in Chapter 6. Table 3.7 shows the examples of the dataset.

Table 3.7: Examples of the WTFC data

t	i	j	k	v	uv	GL	t-t	p-r
2001	4	12	090920	2.321	250.90	0	OWT	L
2001	4	20	570110	1.536	NA	0	OWT	
2001	4	40	490199	0.099	NA	0	OWT	
2001	4	208	430130	2183.260	109711.55	0	OWT	H

※ The column head of Table 3.7 and descriptions are as follows.

t: Year

I: Exporter (ISO 3-digit country code)

j: Importer (ISO 3-digit country code)

k: Product (HS 6-digit code)

v: Value of the trade flow (in 1000 current USD)

uv: Unit value of the trade flow (in current USD / metric ton)

GL: Grubel-Lloyd index

t-t: Trade type: OWT - One-Way Trade, TWT -Two Way Trade, TWH -Two Way Horizontal trade, TWV - Two Way Vertical trade
p-r: Price range: H- High, M -Medium, L-Low

Chapter 4

Surname-origin classifier

4.1 Introduction

Research that classifies nationality and ethnicity using name data has been carried out in biomedicine [Mat07], sociology [DSL07], [AJC07], demographics [JQ11], [JL17], and marketing [SJW15], [App01]. In the first time, classification method was extremely simple. An extensive type of name structure exists, and a limitation where only the names on lists can be classified by the simple method [AA09] that uses name and ethnic list data without changing them. In recent years, Hidden Markov Models and Decision Tree have categorized ethnic groups by name [AA09], by the Bayesian Approach [JC10], with a SVM-based classification of gender[WL13], and a machine learning method [MP11]. Other study analyzed the origin of surnames in French social groups using the Naïve Bayes method [AM18]. This paper adopts a method that estimates nationality from surnames by constructing a language model with a Recurrent Neural Network (RNN). In this chapter, I explain training data and how to build RNN models. And I show the results of the model's classification accuracy with the test data. In addition, I explain how to interpret the model's classification accuracy through comparison with the classification results of related studies and human judgment results. An explanation of the limitations and characteristics of the model is also added at the end.

4.2 Training dataset

I use the ORBIS dataset provided by Bureau van Dijk's in 2016. It provides information about around 300 million companies across the globe as well as public and private company data. The ORBIS 2016 dataset include about 37 million names and nationalities of corporate executives and individual shareholders in 203 countries. The ORBIS dataset basically contains data as follows: the company name, its ID, the number of executives, their names, IDs, job titles, beginning/retirement dates, salaries, dates and places of birth, nationality, and so on. From this dataset, I separately extracted the names, dates and places of birth, nationality, executives' personal IDs, and company IDs for our research purpose. A sample of the resulting data is shown in Table 1. Names were partially * marked to protect privacy. If some data are missing, such as where they came from, they are marked as NA. I use surname and nationality-labeling data only in this paper, cause surname indicate the family root's information of the person. However, more diverse studies will be possible by association with first name, gender, salary, birthplace, and age in the future. I divided the subset by 80% and 20%, as training and test data.

4.3 Training method and setting

The Neural Network model is a mathematical model of the biological brain. In other words, just as biological neurons receive input values and output them externally, artificial neural network neurons receive input values in each layer and send output values to another layer.

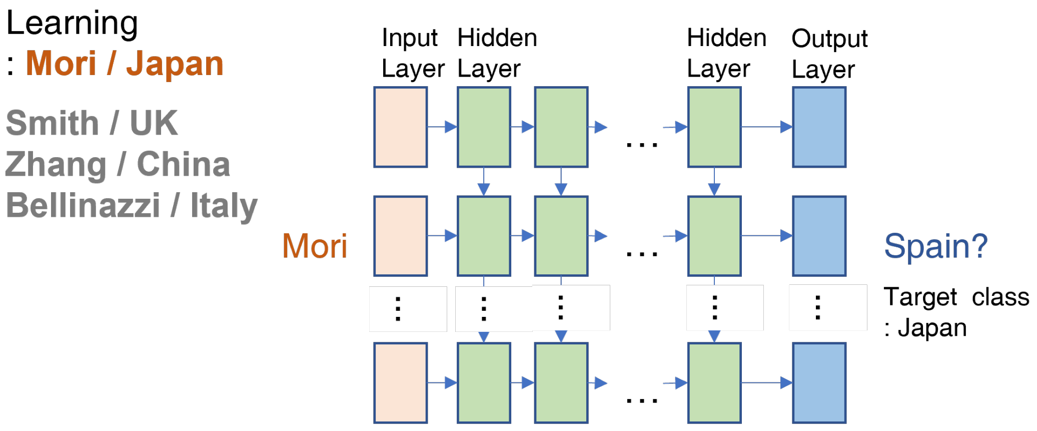


Figure 4.1: RNN training

RNN is a neural network in which hidden layers form a directed graph. Therefore, it can be used for learning continuous data such as text and speech recognition. To use as the training dataset, I decomposed surname into alphabetic characters. For example, "Mori" is decomposed into consecutive alphabet letters: "M" + "o" + "r" + "i". Through such learning, the RNN learns the probability of the character string generation in the surname, such as "o" after "M", "i" after "r", and "r" after "Mo". It also memorizes that this generation probability differs for each nationality by linking it to a surname and learning the nationality together. After the training, if we input surname to the classifier, it returns the probability of the nationality of the surname. With this kind of machine learning method, we can predict its nationality even for surnames that are not in the training dataset. The learning efficiency is changed depending on the learning rate, the number of hidden layers, and the loss function. To solve the Vanishing Gradient Problem, where the RNN's learning ability is degraded by the increase in the distance of the information, we add a Long Short-Term Memory (LSTM) [ref] layer to the RNN. In this research, we test and improve learning efficiency by stacking LSTM layers. I select the top 77 countries with enough data to train. RNN trained the nationalities and surnames of 37,081,935 people in 77 countries. The loss function is cross-entropy (1), where:

$$L(y, \hat{y}) = -\sum_{i=1}^{N} y_i log \hat{y}_i. \tag{4.1}$$

In cross-entropy formula (1), y is the actual nationality distribution obtained from the training data, and y is the nationality distribution estimated from the classifier. N is the number of nationalities (N=77). The learning number is setting to 2 million iterations. Figure 4.2 shows the relationship between the number of learnings in the RNN and the loss value. The drop of the loss value becomes gradual and learning is sufficiently completed when the number of learnings exceeds 400,000 times (= 20 × 20,000 times).

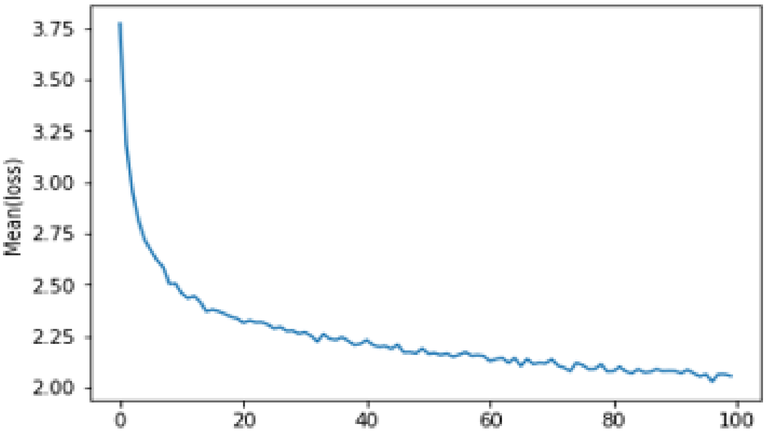


Figure 4.2: Change of losses during RNN learning. The x-axis and the y-axis represent the number of learnings and mean of loss values, respectively.

4.4 Test result

After finish training, a classifier returns log-likelihood calculated from the generation probability of input surnames in each nationality. For example, if a Japanese surname like 'Mori' is input, the log-likelihood shows the highest value in Japan (nationality). In some cases, classification results may be difficult to determine. If some British surname is input, the log-likelihood shows similar values in Australia, New Zealand, and the United Kingdom. Since Australia and New Zealand have British origins in their language, culture, and ethnicity, classifying the nationality of these countries by surname is difficult. To improve the classifier's precision, the log-likelihood values are divided by the number of entries (people) in each countries' dataset and normalized. In this way, our classifier shows a higher score to the United Kingdom than New Zealand if we input the surname like Smith, which is popular in several countries.

Table 4.1: Prediction results.The (number) is log-likelihood of predicted country.

Input	Output (Top3)
Smith	(-0.264) United Kingdom (-0.320) Australia (-0.401) New Zealand
Obama	(-0.246) Kenya (-0.424) Nigeria (-0.470) Japan
Mori	(-0.042) Japan (-0.614) Papua New Guinea (-0.656) Italy

I test the accuracy of the classifier using a randomly selected test dataset. Figure 4.3 shows a confusion matrix obtained by classifying test the surnames of 77 countries. The brighter the color is, the higher is the predicted accuracy. For example, Bulgaria, Iceland, Japan, and Thailand are countries with a high percentage of matches to predicted and actual nationalities. On the other hand, Canada, Brazil, the United States, and the United Kingdom show meager matching rates.

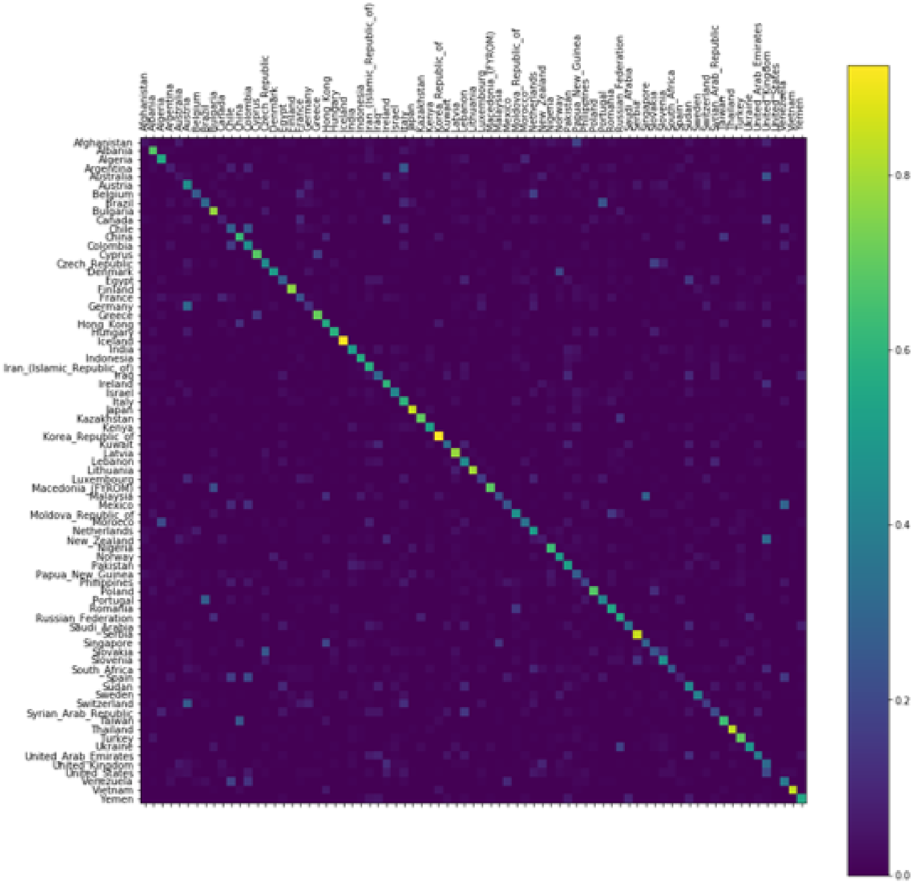
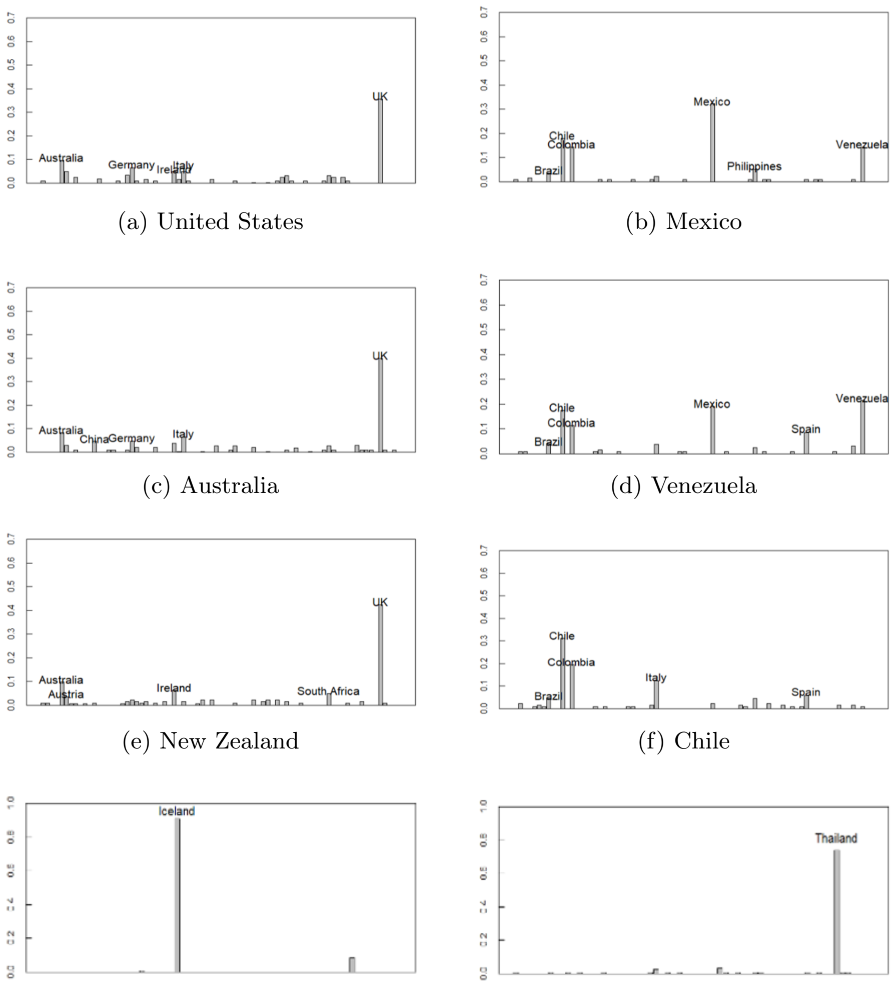


Figure 4.3: Confusion matrix of 77-country. The brighter the color, the higher the accuracy



(a) United States

(b) Mexico

(c) Australia

(d) Venezuela

(e) New Zealand

(f) Chile

(g) Thailand

(h) Iceland

Figure 4.4: Surname-origin(nationality) classification test result in each country. The x-axis and the y-axis represent each country and rate of predicted result, respectively.

Table 4.1 shows the normalized log-likelihood values of the top 3 for each surname. Smith is the most likely to be predicted as the United Kingdom. The surname Mori is clearly much more likely to be predicted as a Japanese surname than in other countries. The log-likelihood value of 'Japan' is high than the others.

Table 4.2: Prediction precision of each country

Rank	Country	Precision (%)
1	Iceland	0.920
2	Republic of Korea	0.901
3	Vietnam	0.883
4	Japan	0.874
5	Serbia	0.822
6	Thailand	0.787
7	Lithuania	0.770
8	Latvia	0.759
9	Bulgaria	0.755
10	Republic of North Macedonia	0.726
11	Finland	0.710
12	Poland	0.697
13	Italy	0.690
14	Turkey	0.672
15	China	0.672
16	Albania	0.656
17	Greece	0.653
18	Cyprus	0.652
19	Pakistan	0.634
20	Taiwan	0.632
21	Hungary	0.623
22	Nigeria	0.614
23	Kazakhstan	0.607
24	Russian Federation	0.605
25	Algeria	0.604
26	Romania	0.563
27	Norway	0.563
28	India	0.549
29	United Kingdom	0.544
30	Iran (Islamic Republic of)	0.532
31	Hong Kong	0.526
32	Kenya	0.505

33	Denmark	0.502
34	Yemen	0.497
35	Ireland	0.488
36	Lebanon	0.480
37	Slovenia	0.471
38	Colombia	0.461
39	United Arab Emirates	0.457
40	Austria	0.454
41	Netherlands	0.454
42	Czech Republic	0.453
43	Indonesia	0.4267
44	Ukraine	0.417
45	Sweden	0.417
46	Moldova Republic of	0.409
47	Israel	0.402
48	Portugal	0.393
49	Saudi Arabia	0.374
50	Malaysia	0.368
51	Chile	0.366
52	Mexico	0.362
53	Brazil	0.358
54	Kuwait	0.338
55	Germany	0.333
56	France	0.316
57	Sudan	0.303
58	South Africa	0.299
59	Papua New Guinea	0.280
60	Morocco	0.273
61	Egypt	0.269
62	Belgium	0.235
63	Switzerland	0.234
64	Iraq	0.229
65	Venezuela	0.225
66	Slovakia	0.212
67	Philippines	0.170

68	Spain	0.165
69	Singapore	0.136
70	Syrian Arab Republic	0.113
71	Luxembourg	0.093
72	Australia	0.087
73	Afghanistan	0.026
74	Canada	0.011
75	Argentina	0.005
76	United States	0.003
77	New Zealand	0.002

Table 4.2 shows the prediction precision of each country. The high-ranking countries are Iceland (0.920), Republic of Korea (0.901), and Vietnam (0.883). The low-ranking countries are New Zealand (0.002), United States (0.003), and Argentina (0.005).

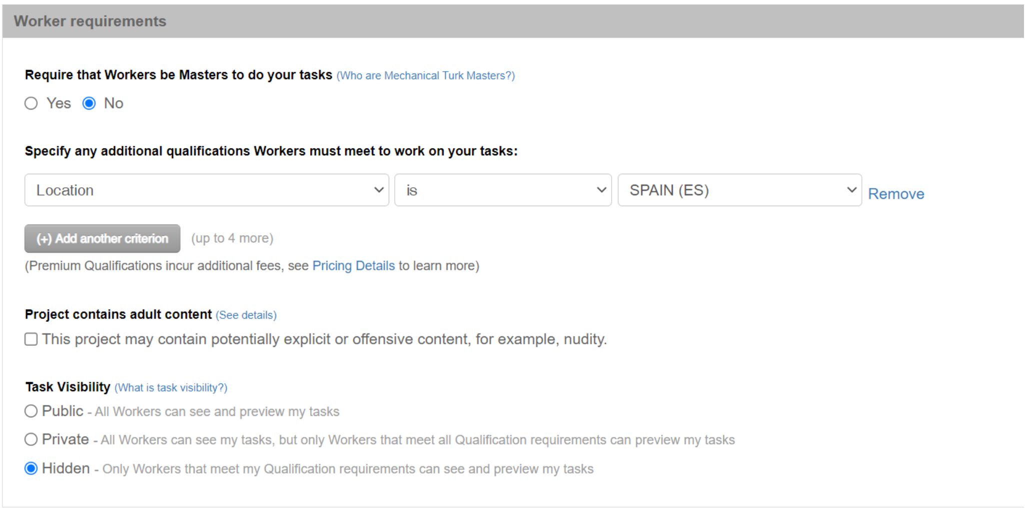
The Classifiers are generally constructed by neural networks that train general features and ignore irregular ones to raise the classification precision. Rather than guess about rare surnames (i.e., Schmidt in China), it is more efficient to match nationality to frequent surnames accurately. In other words, the nationality predicted by the classifier indicates which nationality's feature is most reflected. For example, if the surname of an Italian immigrant of American nationality is input, the normalized log-likelihood shows that its maximum value is in Italy. Using this feature, I estimate the distribution of common nationality in each country. The result, the nationality came out from the classifier, will be called '(surname) origin' in my work.

Figure 4.4 shows the countries where the distribution of the predicted origin is mainly scattered. In the United States, which has many immigrants, the countries of the origins of many people are the United Kingdom, Australia, Germany, Italy, or Ireland. The same tendency is observed in Australia and New Zealand, which also have many immigrants from the United Kingdom. Geographically, Mexico is closer to the United States, but its origin composition is more intimate to Venezuela than the United States. Chile is a little different from Venezuela and Mexico; it shows a high distribution of Italian descent.

4.5 Classification performance

4.5.1 Human judgement

Since I built the classifier with a dataset of surnames and nationalities, its classification is based on nationality. Nationality classification groups contain noisy data than origin groups. However, I argue that the overall characteristics of the group are based on language and culture. Therefore, it can be similarly classified as surname data that reflect it. To support this argument, I compared data classified by origins by human judgment results. I used the Amazon Mechanical Turk [Ama] to collect human judgment data. I requested to the registered people in each country classify the selected sample data. Sample data were randomly extracted from the ORBIS dataset. The request is simple. People in each region will see a list of surnames and then select all the surnames which are originated from the country. For example, a person who lives in Italy see surnames from the ORBIS dataset registered as Italy. And he/she selects all the surnames from the list that he/she believes Italian origin. I put options in the surname list to filter out a person's classification results by reliability. Several of the surnames in the list precisely originated in that region (based on Wikipedia data), and any work that did not select these surnames correctly was filtered out. I only selected reliable results through the answering time. Only those surnames chosen by more than 90% of the people in a region were classified as correct. (I will call it MTurk dataset after now). I extract the data from six countries, the United Kingdom, France, Germany, Italy, India, and Spain, where have sufficient data for analysis through data cleaning. We compared the classified results using the same sample with my classifier. The classification results of the MTurk dataset using my classifier are shown in Table 4.3. The average of the top 1 accuracy is 67%.



Worker requirements

Require that Workers be Masters to do your tasks (Who are Mechanical Turk Masters?)

○ Yes ● No

Specify any additional qualifications Workers must meet to work on your tasks:

Location | is | SPAIN (ES) | Remove

(up to 4 more)

(Premium Qualifications incur additional fees, see Pricing Details to learn more)

Project contains adult content (See details)

□ This project may contain potentially explicit or offensive content, for example, nudity.

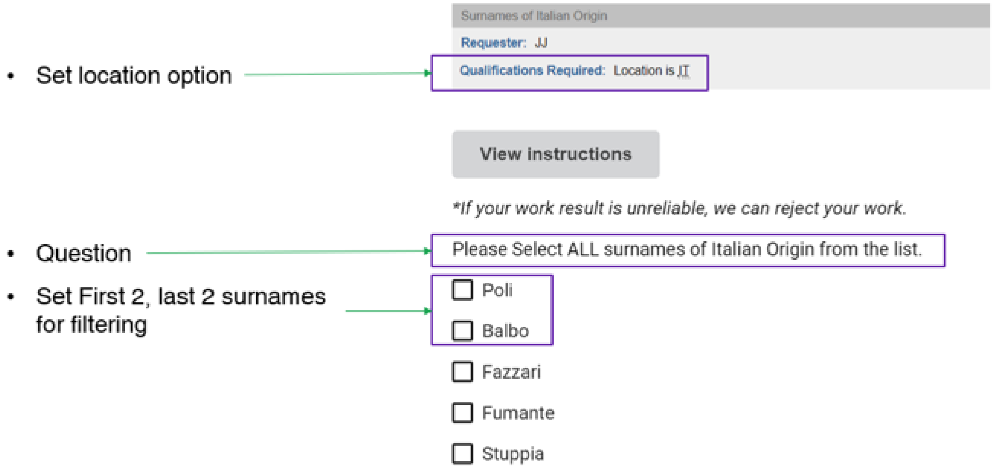
Task Visibility (What is task visibility?)

○ Public - All Workers can see and preview my tasks

○ Private - All Workers can see my tasks, but only Workers that meet all Qualification requirements can preview my tasks

● Hidden - Only Workers that meet my Qualification requirements can see and preview my tasks

Figure 4.5: MTurk setting



- Set location option
- Question
- Set First 2, last 2 surnames for filtering

Surnames of Italian Origin

Requester: JJ

Qualifications Required: Location is IT

*If your work result is unreliable, we can reject your work.

Please Select ALL surnames of Italian Origin from the list.

□ Poli

□ Balbo

□ Fazzari

□ Fumante

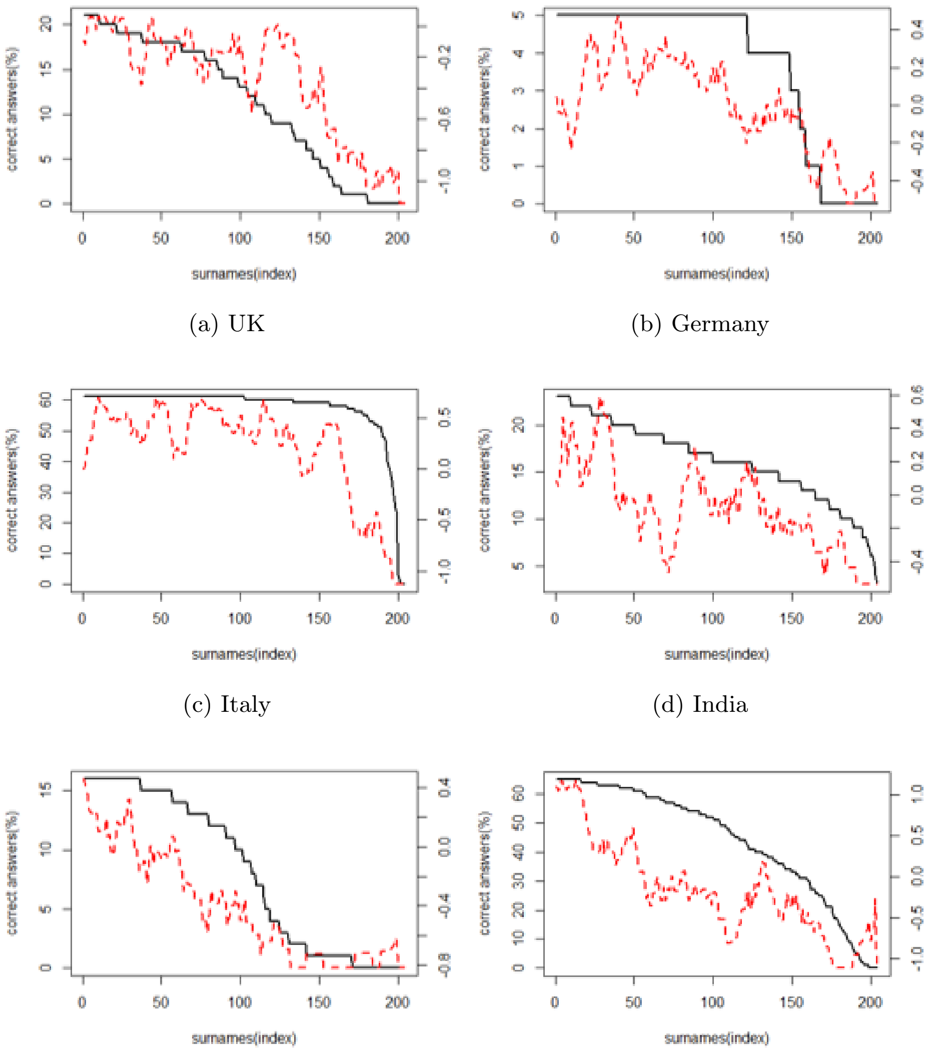
□ Stuppia

Figure 4.6: Request form of MTurk

Table 4.3: Prediction results with MTurk dataset

Region	Balanced accuracy
UK	0.61
Germany	0.68
Italy	0.70
India	0.57
France	0.73
Spain	0.71

It is not a simple task to compare the classification result with MTurk data, because the human judged result is not clearly true. Therefore, I decided to compare the difficulty of classifying ethnicity. If there is a surname which get low answer rate as the originated surname from the region, the surname is difficult to classify the ethnicity by humans. In Figure 4.7, the x-axis shows each surname of the sample. The y-axis (left) represents the ratio of correct answers. The black line represents the distribution of the correct percentage of each surname classified by humans. This indicates how difficult it is for each surname to be correctly classified. For comparison, I use the log-likelihood values from the classifier. And I calculate the difference of log-likelihood values between the top 3 and the top 1 in the classifier results. The lower the difference, the more difficult classification is. To reduce the noise of this calculation, I used a sliding window in which the average size was ten. The red line indicates the predicted difficulty of the classifier. Figure 4.7 shows that my classifier can classify surname origin through surnames to a similar level of human judgment.



(a) UK

(b) Germany

(c) Italy

(d) India

(e) France

(f) Spain

Figure 4.7: Relation between human and classifier results. The black lines represent the distribution of the correct percentage of each name classified by humans. Surnames are sorted by the distribution. The red lines represent the predicted difficulty of the classifier.

4.5.2 Olympic data

I compare the performances of our classifier with related studies. Just a few open datasets are available due to privacy issues. The data most commonly used to build name-nationality or name-ethnic (racial) classifiers in informatics are the names and nationality data found in Wikipedia and Olympic participants data. I can collect these two forms of data by crawling. We compared the performances using the Olympic participants' data, which were published in Kaggle [Gri18].

The Kaggle's data contained 120 years of Olympic athlete data from 1896 to 2016. Out of 271,226 entities from 230 countries, 117,059 unique objects from 77 countries were extracted. The data from these Olympic participants were used as test data to identify the performance of our classifiers. Table**??** shows the balanced accuracy of the predicted results in each country. The average predicted precision of the 77 countries is 0.66, which is higher than the related research [LKK+17] that tested with data from identical Olympic athletes. From the related research [LKK+17], 52.6% of the Top 1 accuracy was predicted in 44 countries by surname-nationality prediction.

This related study [LKK+17] used machine learning with an STML layer to an RNN. This is similar to our classifier model. But the difference in the quality and the amount of the training data probably reduced the precision. For the Olympic participant data, we predicted that the volume of the data would be heavily skewed by the country. Also, many athletes have changed their nationalities, and it makes lowering the performance. On the other hand, compared to the number of Olympic participants, using highly reliable training data allowed for better performance than the related research [LKK+17].

Table 4.4: Balanced accuracy of Olympic participant's dataset prediction

Rank	Country	Balanced accuracy
1	Iceland	0.946
2	Japan	0.910
3	Bulgaria	0.864
4	Greece	0.845
5	Thailand	0.832
6	Macedonia (FYROM)	0.818
7	Algeria	0.807
8	Finland	0.795

9	Albania	0.793
10	Nigeria	0.791
11	Denmark	0.790
12	Lithuania	0.789
13	Latvia	0.783
14	Ireland	0.781
15	Sweden	0.777
16	Poland	0.774
17	Portugal	0.772
18	Italy	0.760
19	Hong Kong	0.759
20	Pakistan	0.755
21	Iran (Islamic Republic of)	0.732
22	Vietnam	0.730
23	Indonesia	0.728
24	United Kingdom	0.717
25	Russian Federation	0.713
26	Ukraine	0.712
27	Hungary	0.709
28	Romania	0.707
29	Cyprus	0.707
30	India	0.706
31	Netherlands	0.703
32	Slovenia	0.693
33	Moldova Republic of	0.687
34	Norway	0.685
35	China	0.676
36	Turkey	0.674
37	France	0.667
38	Austria	0.655
39	Belgium	0.651
40	Germany	0.649
41	Kenya	0.641
42	Colombia	0.633
43	Iraq	0.625

44	Egypt	0.623
45	Switzerland	0.618
46	Philippines	0.613
47	Taiwan	0.613
48	Serbia	0.611
49	Spain	0.610
50	Kazakhstan	0.610
51	Lebanon	0.606
52	Papua New Guinea	0.606
53	Luxembourg	0.605
54	Brazil	0.603
55	Yemen	0.602
56	Syrian Arab Republic	0.594
57	Korea Republic of	0.593
58	Morocco	0.592
59	Israel	0.591
60	Sudan	0.589
61	Malaysia	0.588
62	Chile	0.584
63	Mexico	0.583
64	South Africa	0.580
65	Czech Republic	0.578
66	United Arab Emirates	0.565
67	Singapore	0.560
68	Saudi Arabia	0.538
69	Slovakia	0.538
70	Venezuela	0.533
71	Argentina	0.527
72	Kuwait	0.522
73	United States	0.517
74	New Zealand	0.510
75	Australia	0.506
76	Canada	0.503
77	Afghanistan	0.497

	Average	0.668

The test and comparison results show that this classifier does not classify nationality by surname. Using this classifier, Zheng, Yamato, and Muller, who lives in the United States, are not classified as American nationalities but as Chinese, Japanese, and German origins, respectively. As MTurk data test results show, we can see that the classification is similar to how humans classify the origin of a surname. Since this surname-origin classifier derives classification results only with Top1, the classification accuracy is low in the situations shown in the table 4.5 below. In other words, Origin, which is in a similar language area and uses many similar surnames, has a technical limitation: the accuracy is lowered. To solve this problem, it is thought that one way to present classification candidates from Top1 to Top5 or to adopt only the classification result with a significant difference between Top1 and Top2.

Table 4.5: Cases that cause classification errors

Surname	Correct Answer	Classified (log-likelihood)
Lorenzen	German	’UK’(-0.3068), ’Spain’(-0.3359), ’France’(-0.3598)
Meier	German	’France’(-0.2461), ’Germany’(-0.2793), ’Spain’(-0.3672)
Prieto	Spain	’Italy’(-0.0752), ’Spain’(-0.2763), ’France’, (-0.5048)
Ahijado	Spain	’Japan’(-0.1332), ’Spain’(-0.4548), ’Israel’(-0.5149)

4.6 Summary

In this chapter, I explained how to build a surname-origin classifier that identifies 77 origins using ORBIS’ large-scale surname-nationality labeling data and RNN. In addition, to evaluate the performance of the classifier, the accuracy evaluation through test data, using Olympic data used in related studies, and evaluation results compared with MTurk’s human judgment data are also shown. The surname-origin classifier constructed in this study can identify the origin of a surname similarly to the level judged by humans. However, accuracy is different depending on the origin, and there is room for improvement by adopting the difference between Top1 to Top2 results or showing the groups of Top1 to Top5 results.

Chapter 5

Application to sociology

5.1 Introduction: diversity in sociology

This chapter describes applied sociological research, especially on diversity, using the surname-origin classifier constructed in Chapter 4. Diversity is the range of differences between people. Diversity includes various attributes such as sex, gender, race, ethnicity, age, social class, and ability. In this study, I focus on diversity according to the origin of surnames. It is analyzed by dividing into countries, social groups, and regions.

5.2 National origin diversity

Each country has different ethnic diversity. Some countries have an overwhelmingly high proportion of the dominant ethnic group, while others have a diverse ethnic population. Ethnicity classification has different standards of detail for each country, but comparative analysis between nations is possible through the surname-origin classifier.

Japan's 98.5% of the population is Japanese. In Kenya, which is a multi-ethnic nation, the dominant ethnic group (Kikuyu) accounts for only 17%. I measure ethnic diversity by estimating ethnic distribution from the relationship between the entropy (5.1) of the ethnic distribution of each country and the proportion of the dominant ethnic group in each country [lana], [lanb], [fb]. Figure5.1 shows the relationship: the higher the entropy, the greater the ethnic diversity (multi-ethnic nation). As such, the entropy of the ethnic distribution through surname data can predict a country's ethnic diversity.

$$H(X) = -\sum_{i=1}^{N} p(x_i) log_2 p(x_i). \tag{5.1}$$

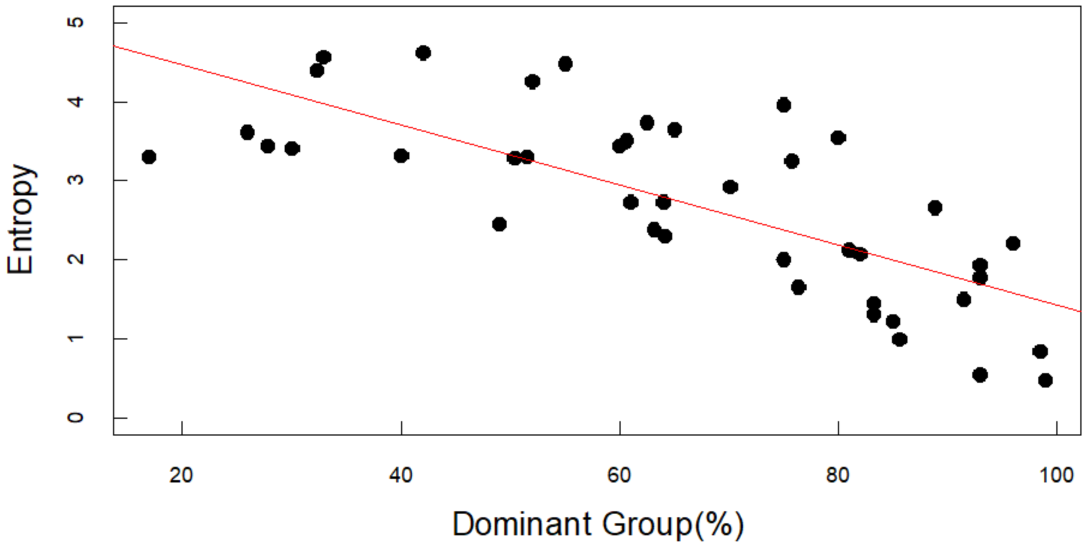


Figure 5.1: Relationship of entropy and dominant group. The X-axis indicates the rate of the dominant group in the country's population. The y-axis shows the information entropy of the predicted result (predicted distribution of each country)

Table 5.1: Entropy and dominant group of each country

Rank	Country	Dominant Ethnic Group	%	Entropy
1	Republic of Korea	Korean	99	0.485
2	Iceland	Icelandic	93	0.561
3	Japan	Japanese	98.5	0.853
4	Vietnam	Vietnamese	85.7	1.005
5	Bulgaria	Bulgarian	85	1.224
6	Serbia	Serbs	83.3	1.306
7	Lithuania	Lithuanian	83.3	1.444
8	China	han	91.5	1.504
9	Thailand	Thai	76.4	1.661
10	Finland	Finnish	93	1.780
11	Greece	Greek	93	1.933
12	Turkey	Turkish	75	2.011
13	Ireland	Irish	82	2.073
14	Russian Federation	Russian	81	2.132
15	Italy	Italian	96	2.214
16	Macedonia (FYROM)	Macedonians	64.2	2.296
17	Kazakhstan	Kazakhs	63.1	2.386
18	Colombia	Mestizo	49	2.458
19	Romania	Romanians	88.9	2.655
20	Czech Republic	Czech	64	2.740
21	Iran (Islamic Republic of)	Persian	61	2.740
22	Lebanon	Lebanese	70.2	2.922
23	Moldova Republic of	Moldovans	75.8	3.259
24	Malaysia	Malay	50.4	3.287
25	Kenya	Kikuyu	17	3.294
26	Venezuela	Mestizo	51.6	3.303
27	Indonesia	Javanese	40	3.320
28	Chile	Caucasian	30	3.401
29	Kuwait	Arab	60	3.441
30	United Arab Emirates	Indian	27.8	3.444
31	Saudi Arabia	Saudiarabian	60.7	3.507
32	Germany	Germans	80	3.545
33	Australia	English	25.9	3.610
34	Switzerland	Germans	65	3.651
35	Argentina	Italian	62.5	3.739
36	Israel	Jewish	75	3.961
37	Belgium	Flemish	52	4.255
38	Canada	Canadian	32.3	4.388
39	Luxembourg	Luxembourgers	55	4.486
40	Philippines	Visayan	32.9	4.563
41	Afghanistan	Pashtun	42.1	4.621

5.3 Organizational origin diversity

If surname data for a specific social group exists, it is possible to analyze the characteristics of the social group. I analyzed the racial composition of the American Physician using a list of Physician donators during the 2012 Federal Selection in the United States. At this time, according to the US census data, the estimated ethnic groups were largely grouped into four races and compared with other social groups. I compare the race composition of public figures with the ORBIS data and the race composition of public figures with the Dow Jones Watchlist data [Dow]. Table 5.2 shows the racial composition of the United States, and table 5.3 shows the proportion of races for each social group. Figure 5.2 shows how different the ethnic proportions of each social group differ based on Census population statistics. It can be seen that the proportion of whites is high in the public figure group, and that the proportion of Asian descent is higher than that of the population statistics in the business and physician groups. In the case of Hispanics, the figure is low in all social groups, indicating that the social advancement of Hispanic groups differs from other races.

Table 5.2: Racial composition of the United States

Race	Origin (country)
Non-Hispanic whites	France, Germany, Ireland, Italy, Poland, UK
Hispanics, Latino	Colombia, Mexico, Spain, Venezuela, Chile
Black, African American	Nigeria, DR. Congo, Angola, South Africa, Egypt
Asian	China, India, Japan, Rep. of Korea, Vietnam

Table 5.3: The proportion of races for each social group in the United States

Race	Population	prediction	Business	Watch List	Physicians
Non-Hispanic whites	69.13 %	68.8%	69.9%	75.7%	61.5%
Hispanics	12.5%	11.3%	6.9%	8.2%	11.0%
African Americans	12.0%	13.1%	12.3%	10.2%	13.9%
Asian Americans	3.6%	6.6%	10.8%	5.7%	13.5%

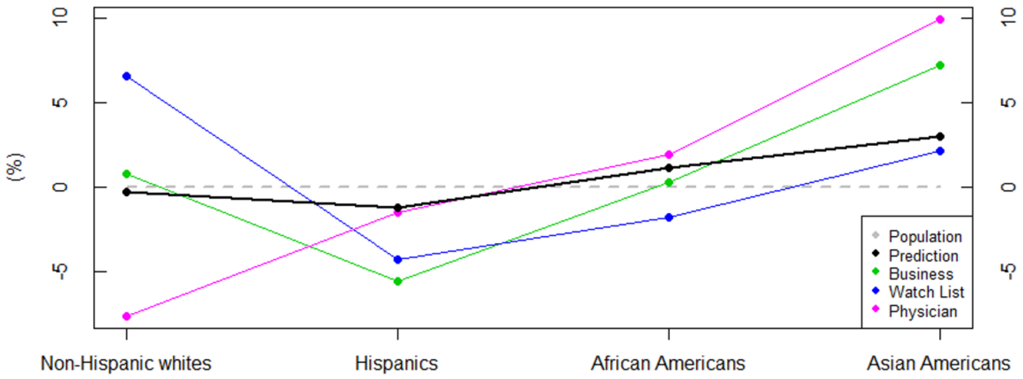


Figure 5.2: Comparison of the percentage of races in the entire population with the percentage of races in each social group

5.4 Regional origin diversity

Using each country's origin distribution, countries with similar ethnic compositions can be extracted. I can trace the connectivity of countries by measuring the similarity among the origin distributions obtained through my classifier and extracting the communities estimating the similarity. I use Jensen-Shannon Divergence (JSD) for calculate the similarity of the distribution. It uses Kullback-Leibler Divergence (KLD) to calculate the distance among distributions. KLD 5.2 calculates the information loss rate between the probability distribution and approximate probability distribution. Because of KLDs have asymmetry, it is unsuitable for calculating distances. Thus, I need the mean of the two probability distributions and the mean of the KLD between each distribution. From this process, I can obtain symmetrical figures: JSD 5.3. And it can be interpreted as the distance between the two probability distributions.

$$D_{KL} = (P||Q) = \int_x P(x)(logP(x) - logQ(x)) \tag{5.2}$$

$$D_{JS} = \frac{1}{2}(P||\frac{P+Q}{2}) + \frac{1}{2}D_{KL}(Q||\frac{P+Q}{2}) \tag{5.3}$$

And, using the similarity of the origin distribution calculated by JSD, I extract the communities of regions (countries or cities) with similar ethnic compositions. There are Various algorithms can detect network communities. In this study, I use the Map Equation [RAB09]. Map Equation, which is based on modularity maximization, extracts communities by considering patterns of network structure. Based on entropy measurements, finding the partitions in a network that can reduce the movement of a random walker becomes the algorithm's core.

The African continent is a very diverse blend of ethnicities, languages, and religions. But due to the lack of accurate demographic data, it is difficult to understand the composition of society. My classifier trains surname and nationality data of 48 countries of the African continent and estimates their ethnic composition. Then by comparing the estimated origin distribution of each country, the similarities of ethnic composition among the nations can be visualized. By extracting the communities, I can clarify the spatial distribution of origins of African countries.

Figure 5.4 shows the estimated origin distribution of Morocco, Cameroon, and the Central African Republic. Morocco and Cameroon have high diversity. On the other hand, the Central African Republic has a relatively low diversity. Morocco resembles the origin distribution of Algeria, Tunisia, and Egypt because they are geographically close. Cameroon is located in the middle of Africa. The Central African Republic has a similarity of origin composition with neighboring countries such as Gabon, Benin, and Chad. These results related to not only the geographical effect but also the historical context where European colonial powers established African borders.

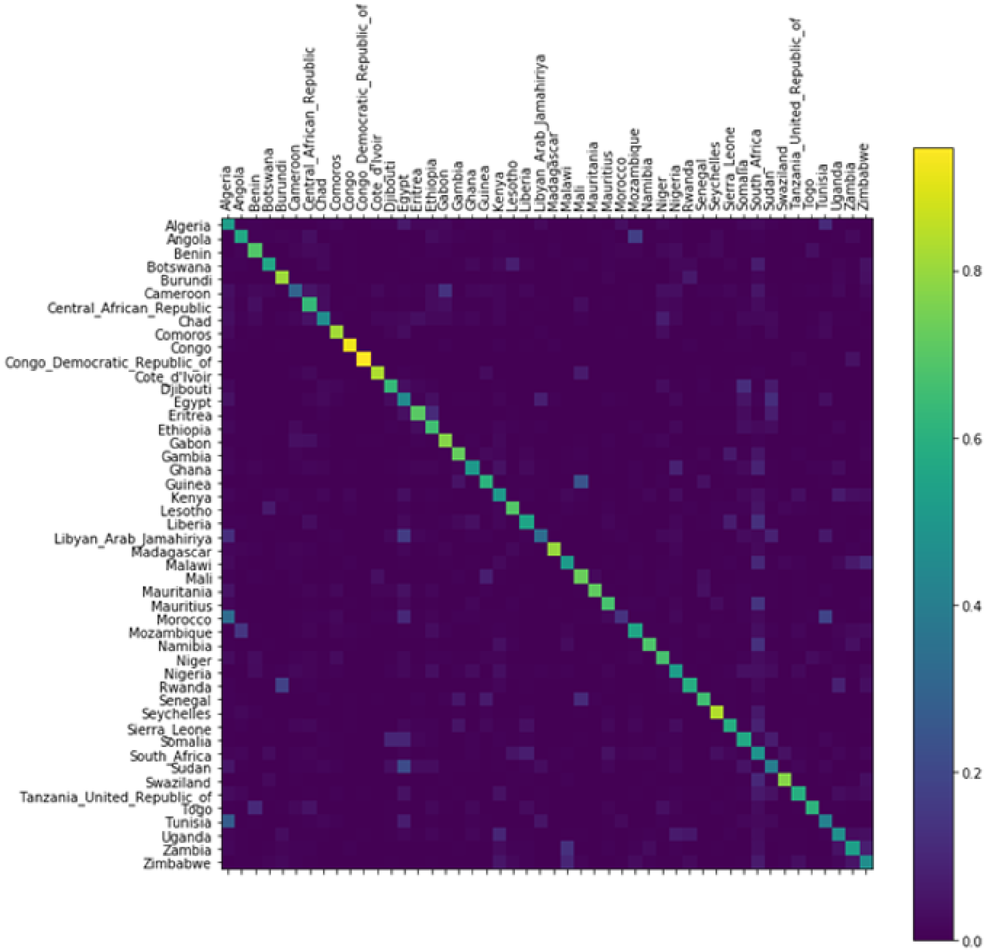
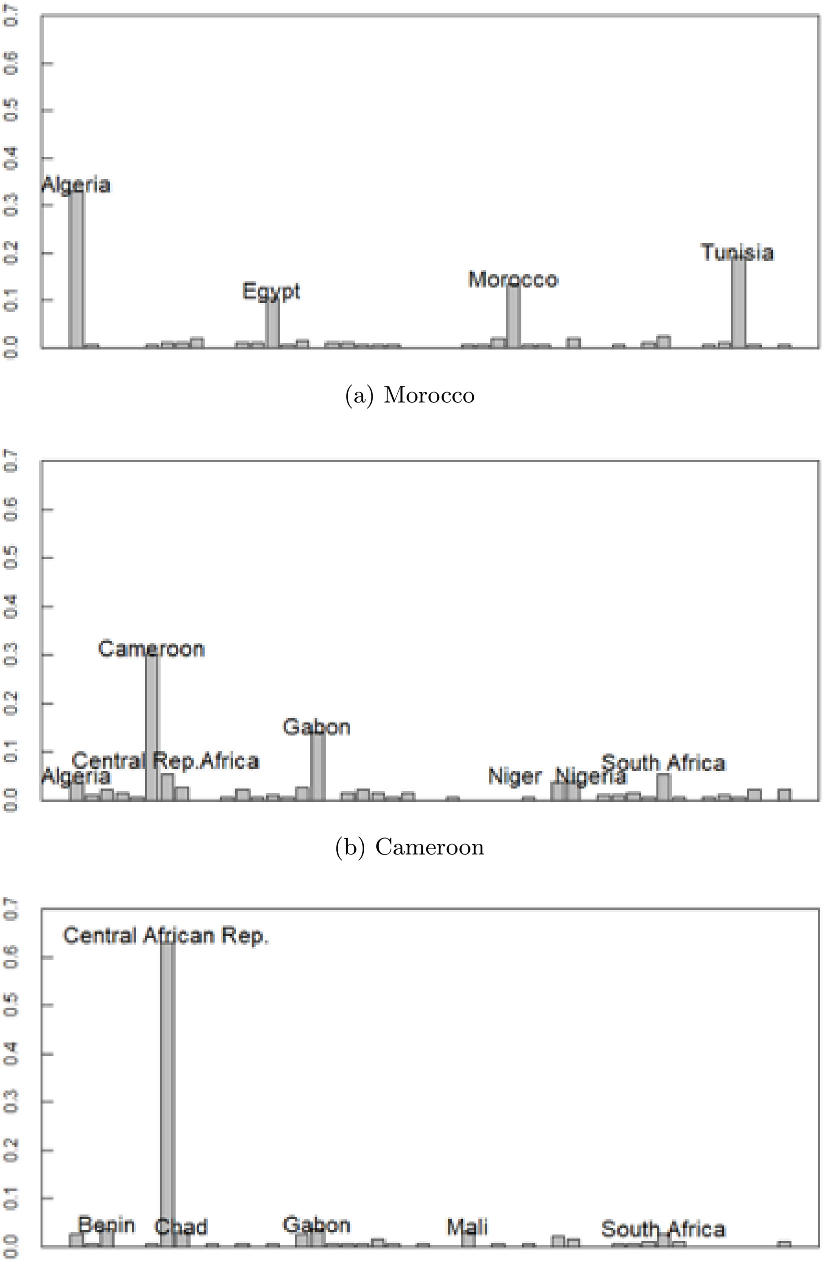


Figure 5.3: Confusion matrix of African continent. The brighter the color, the higher the accuracy.



(a) Morocco

(b) Cameroon

(c) Central African Republic

Figure 5.4: Example of origin distribution in Africa. The x-axis and the y-axis represent each country and rate of predicted result, respectively.

And I quantify the ethnic connections among countries by measuring their similarities with the probability distribution of origins using JSD. Figure 5.5 represents the ethnic network of a community of Africa. The nodes indicate each country, and the link shows the relationship. The thickness of the link is from the calculated value of JSD. I removed links with JSD of 0.09 or less. The countries with the similar origin composition are clustered together. With Map Equation [RAB09], I can detect clusters in a network to capture the spatial connections of origins among countries. The four clusters that we detected were color-coded and mapped to the Africa map of Figure 5.5 White-colored areas represent countries that were excluded from analysis due to insufficient data. Geographically close countries are grouped as a cluster whose shape resembles the spatial distribution of Africa's languages and religions.

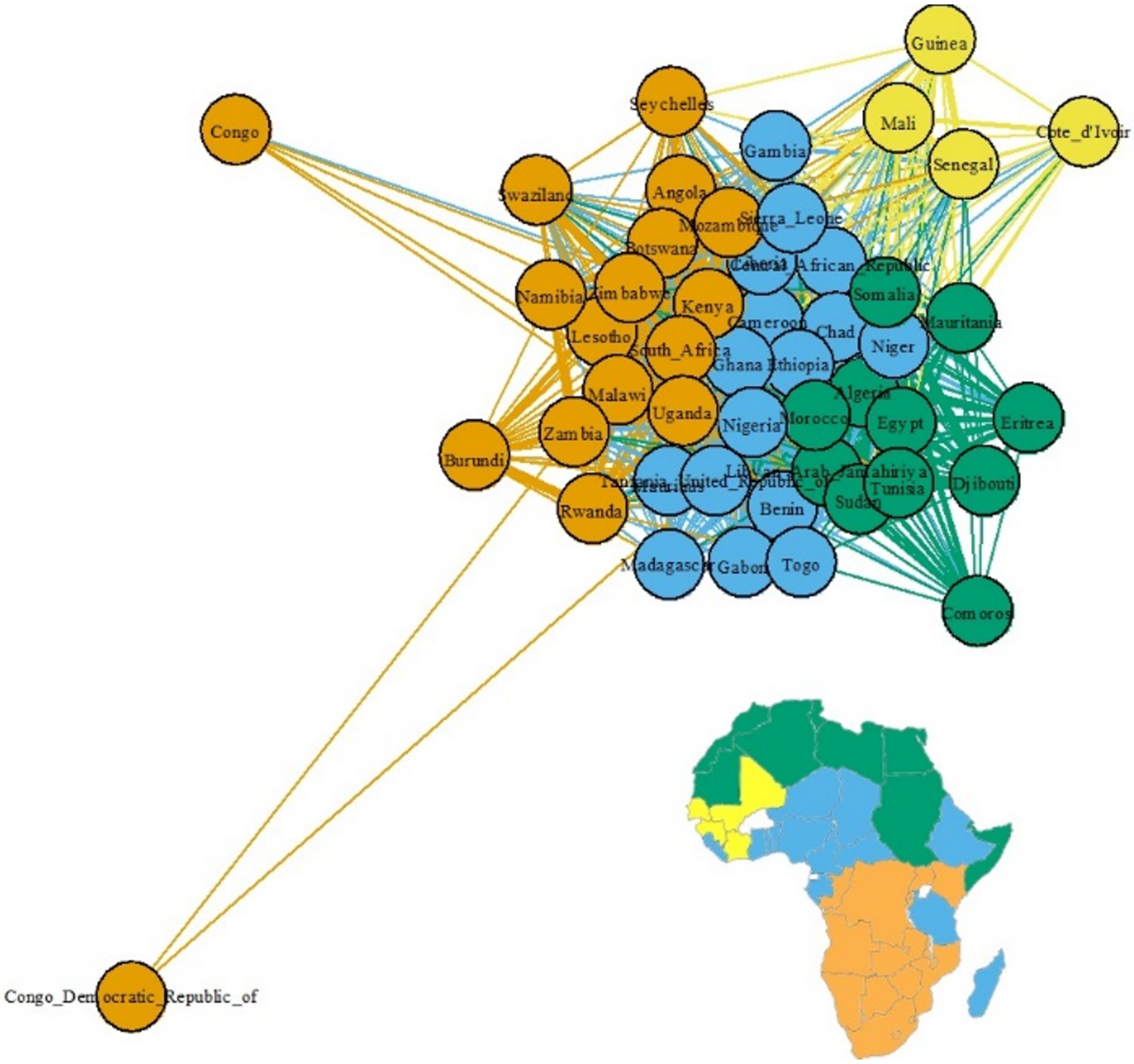


Figure 5.5: Network clustered by similarity of origin composition. The countries are grouped by the similarity of their origin distribution. The same color means the same group and uses the same color in the African map on the lower right.

I analyze European countries in the same way as the analysis of the African continent. Despite the complexity of origin composition in European countries, it is not easy to analyze ethnic groups due to the social atmosphere that does not want to distinguish races separately. However, the method used in this study can confirm the regional ethnic composition. I analyze the origin composition of famous cities in France and Switzerland. Figure5.6 shows the origin distribution of Strasbourg, which is close to France's representative city of Paris, Germany, and Lyon, which is close to Italy. Strasbourg has a higher proportion of German origin than the other two cities. And in the case of Lyon, it can be seen that the proportion of French and Italian is high. Figure5.7, Figure5.8 is the result of clustering each city according to the similarity of origin composition. France is largely divided into two groups. In particular, it can be seen that cities close to Germany show different characteristics in origin composition from the rest of the cities.

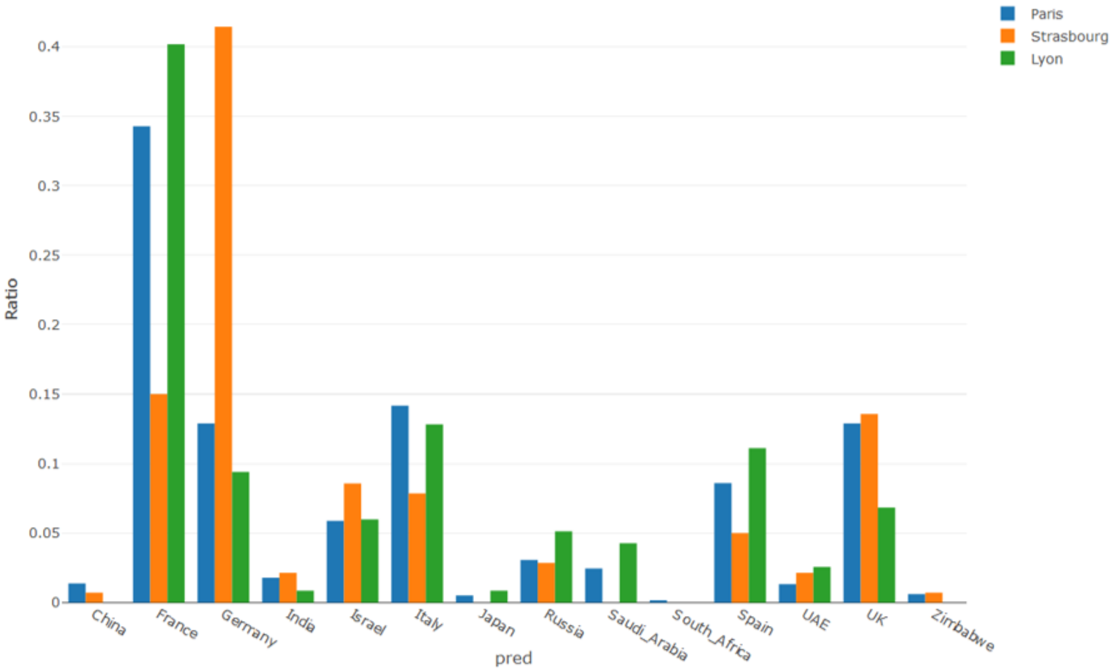


Figure 5.6: Origin distribution of France cities

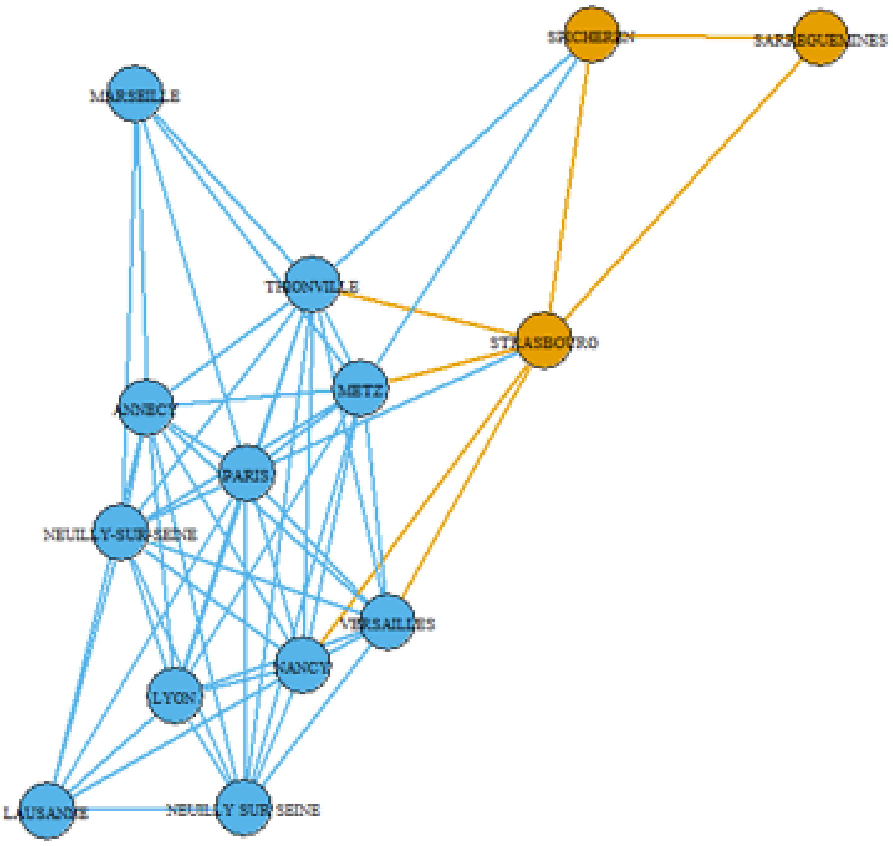


Figure 5.7: Spatial distribution of origins in France(clustering)

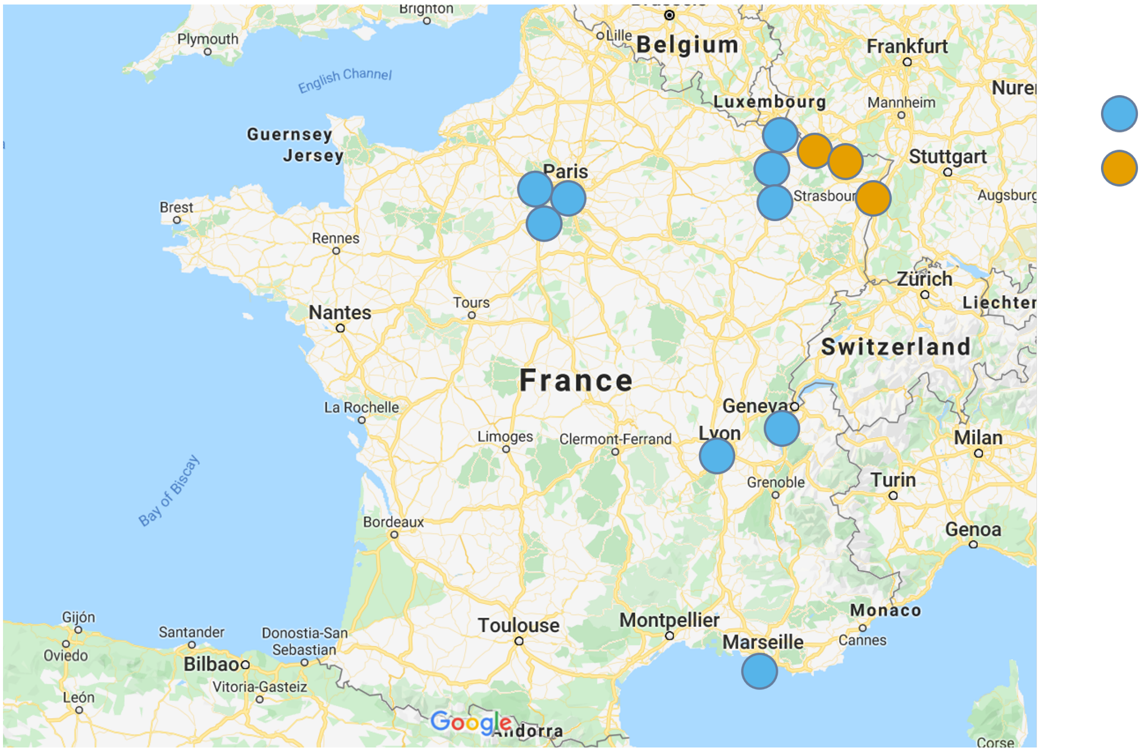


Figure 5.8: Spatial distribution of origins in France(mapping)

Figure5.9 shows the analysis results of Switzerland. It can be seen that the origin composition of the three representative cities of Switzerland is very different origin composition, which is located close to Germany, France, and Italy. Unlike Basel and Zurich, where the composition of Germany is high, Geneva has a high proportion of French and Italian groups. Figure5.10, Figure5.11 shows the results of clustering through the similarity of origin composition in each city. It can be seen that the group was formed largely south and north. Only Frienhach appears to have an independent origin composition. In the case of European countries with borders facing each other, it can be seen that the origin composition of nearby cities varies due to the influx of origins from adjacent countries. Although this study did not analyze the temporal change, it is thought that the movement of origin groups can be observed in the analysis considering time.

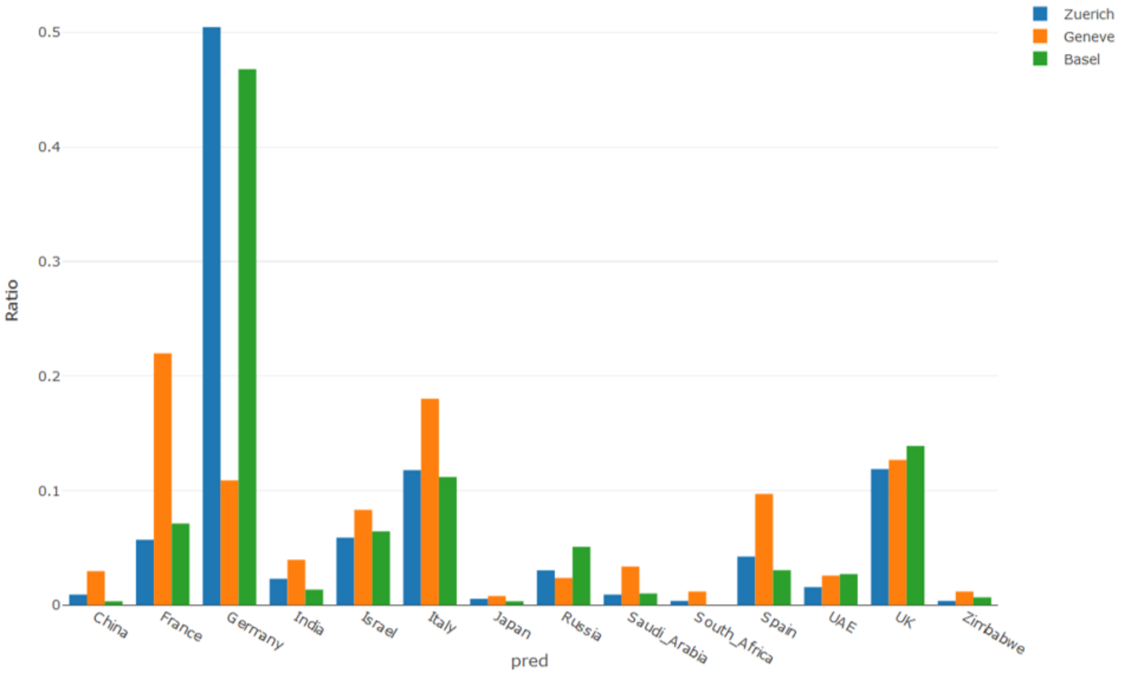


Figure 5.9: Ethnic distribution of Switzerland cities

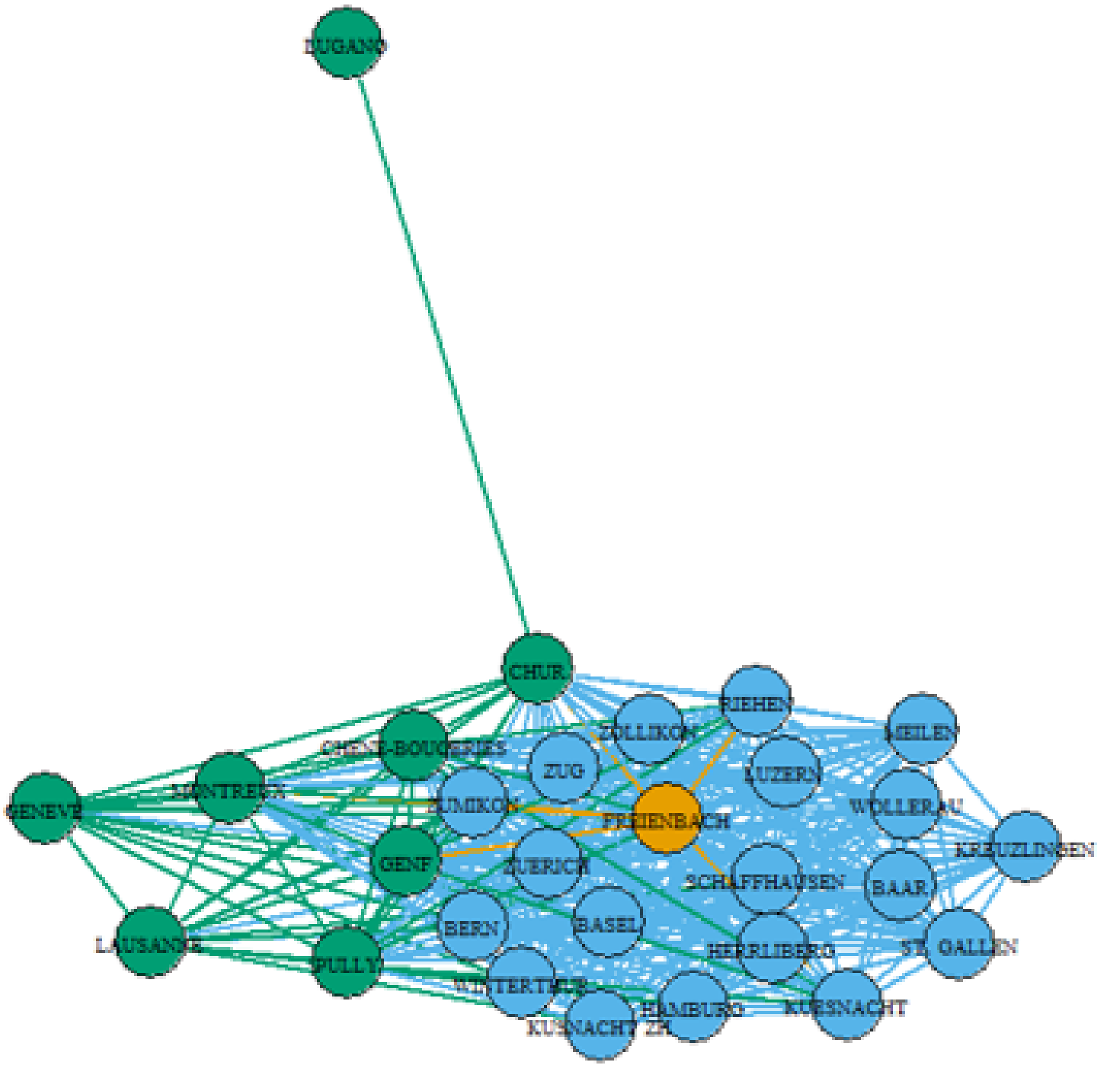


Figure 5.10: Spatial distribution of ethnicity in Switzerland(clustering)

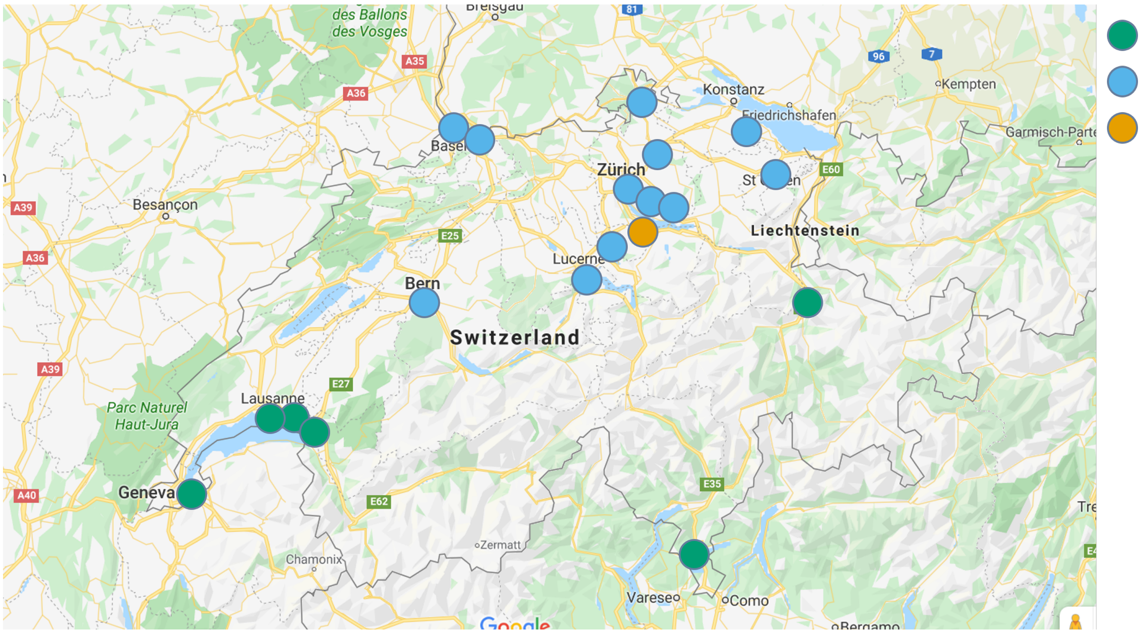


Figure 5.11: Spatial distribution of origin in Switzerland(mapping)

5.5 Summary

This chapter introduced applied sociological research to analyze origin diversity by country, social group, and region. Diversity by country was compared and analyzed by calculating entropy from the origin's distribution classified by surname. The diversity of social groups within a country could also be compared. Based on the distribution of the US population, I showed which races are relatively active in the groups of physicians, public figures, and business people. In addition, the spatial distribution of origins in the region was also visualized. Regional communities with similar origins were extracted through the African continent and European countries. In this study, a comparative analysis was conducted through surname data of the same period. Still, if surname data by time is secured, it is possible to study the movement and change of the origins.

Chapter 6

Application to economics

6.1 Introduction: interethnic trade

Previous studies found that ethnic factors affect cross-country trade, although their scope was limited to a single ethnic group. The reason is that classifying ethnic groups and collecting data are complicated. The ethnic network used by many related studies is based on immigration data. The impact of ethnic networks is analyzed using the statistical data of immigrants who migrated from their birth country to another country [Tri02], [GJF10], [IMZ20], [DW02]. If the composition of ethnic groups is too complex and obtaining statistical data is difficult, their analysis is based on research datasets or survey data [Cop18], [HE14], [Iwa14], [JCA14]. However, it remains precarious since disentangling ethnic groups depends on social environments. Classifying ethnicity is challenging because it is affected by such factors as race, culture, language, and region. Therefore, race and cultural homogeneity are difficult to consider when their classification is based on a single country. Addressing ethnic diversity in multi-ethnic countries is also complicated. To solve this problem, I use surname data to obtain ethnicity. Surnames are influenced by ethnicity, culture, and region. Because of these characteristics, surname data are actively used in research to estimate race and ethnicity [AA09], [JC10], [MP11]. In addition, the ethnic factors estimated by surnames are applied to a wide range of studies, including sociology [JQ11], demographics [DSL00], biomedicine [EGB03], etc. I analyze ethnic homophily in international trade using the surnames of managerial positions in the companies with our surname-origin classifier built in our previous study [15]. The classifier predicts ethnicity based on surnames. With classified results, we investigate a company's majority ethnic groups of managerial positions, calculate the transaction frequency between ethnicities to find ethnic homophily in trade, and measure its statistical significance using the Gravity model.

6.1.1 What is an ethnic group?

The answer to what ethnicity is varies widely. Since this study is based on ethnic classification, we will first consider what criteria can be used to classify ethnic groups from the perspective of classification.

Ethnicity can be classified into lineage, culture, infrastructure, and language. However, the concept of ethnicity is not simple enough to classify it as one standard. In general, ethnic groups are classified by integrating various criteria rather than one. Therefore, according to the field of research, the criteria for classifying ethnic groups according to the object and purpose vary.

Ethnic groups are classified based on name data. The name is inherited through the family tree, so it includes information of bloodline and linguistic and cultural elements. In other words, since the elements of ethnic classification that can be generally used are mixed, a much more universal classification is possible than when classified with only one element.

In this study, first, a classification model is created by learning name-nationality matching data. This classification model identifies the nationalities that use a particular name a lot. But nationality is an administrative classification criterion. Names commonly used in English-speaking countries, such as Smith, are also classified into different nationalities, such as the United Kingdom, the United States, and Australia. A more general classification is needed in this study. Accordingly, a method of grouping nationality into the concept of ethnicity was used in consideration of linguistic factors. When randomly selected name datasets are classified as classifiers, groups of nationalities with similar results are created. For example, Smith would be classified as a country such as the United Kingdom, the United States, and Australia, and a surname such as Zheng would not be classified as a name used in such a country. Through this process, the name-nationality classifier is transformed into a name-ethnic classifier. The name-ethnic classifier eventually classifies countries of the same language. In other words, in this study, countries in the same language are used as the concept of ethnicity. The results of integration into the same language-speaking country were very similar to the categories that classified ethnic groups into language and culture in previous studies. I improved the classification accuracy by clustering 77 countries with a surname vector defined by the average of the log-likelihoods obtained when classifying each surname. I used k-means clustering, where k=17 is determined by the silhouette method [Rou87] (Figure 6.1). These 77 countries are clustered into 35 ethnic groups (Table 6.1). Each ethnicity group is identified by the ISO code of the country with the most business people(managers) who belong to it. Multi-ethnic countries with a similar distribution of surname origins, such as the U.S. and the U.K., are grouped into the same ethnic group. The classification accuracy for ethnic group units is high, e.g., 0.606 for the GB group. I determined a company's majority ethnic group from the surnames of its managers and investigated the impact of ethnicity on international trade between companies.

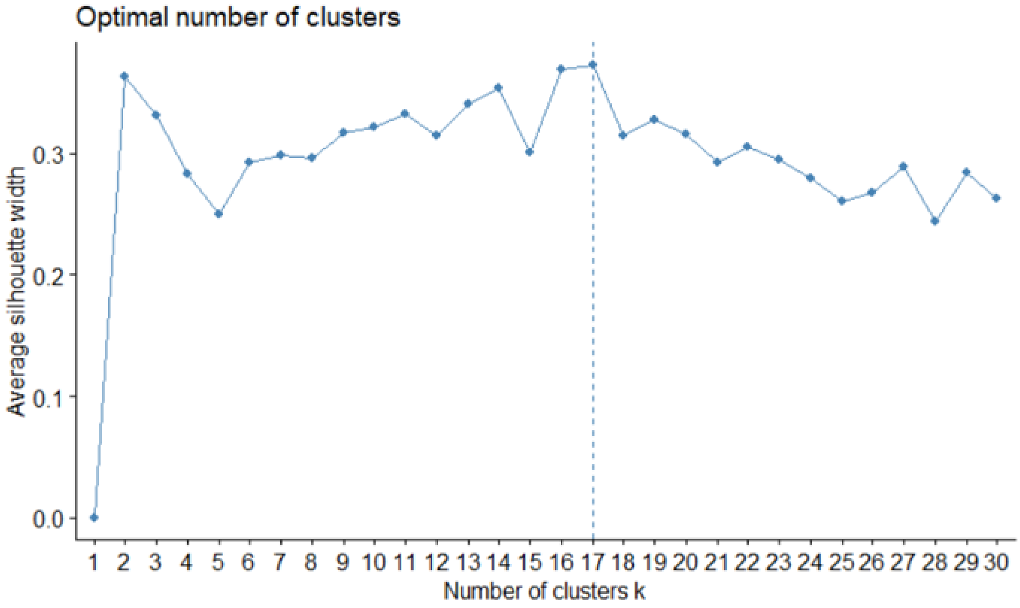


Figure 6.1: Optimal number of Clusters using Silhouette method

Table 6.1: Clustering results of surname origins. Numbers in parentheses refer to classification accuracy. In the FactSet shipping dataset, the upper 16 groups (from GB to TH) account for 97% of the total transaction records.

Ethnic groups	Countries
GB (0.606)	Australia (0.087), Canada (0.010), Ireland (0.488), New Zealand (0.001), South Africa (0.299), United Kingdom (0.544), United States (0.003)
AT (0.456)	Austria (0.454), Belgium (0.235), Netherlands (0.454)
CN (0.675)	China (0.672), Hong Kong (0.526), Malaysia (0.368), Singapore (0.136)
IT (0.690)	Italy (0.690)
DE (0.382)	France (0.316), Germany (0.333), Luxembourg (0.093), Switzerland (0.234)
IN (0.433)	Afghanistan (0.026), India (0.549), Indonesia (0.427), Israel (0.402), Papua New Guinea (0.280)
CO (0.562)	Colombia (0.461), Mexico (0.362), Venezuela (0.225)
BR (0.504)	Argentina (0.005), Brazil (0.358), Chile (0.366), Philippines (0.170), Portugal (0.393), Spain (0.165)
JP (0.874)	Japan (0.874)
SE (0.647)	Denmark (0.501), Norway (0.563), Sweden (0.417)
TW (0.632)	Taiwan (0.632)
AE (0.513)	Iraq (0.229), Kuwait (0.338), Saudi Arabia (0.374), Sudan (0.303), United Arab Emirates (0.457)
EG (0.592)	Egypt (0.269), Islamic Republic of Iran (0.532), Lebanon (0.480), Pakistan (0.634), Syrian Arab Republic (0.113)
KR (0.901)	Republic of Korea (0.901)
TR (0.672)	Turkey (0.672)
TH (0.787)	Thailand (0.787)
UA (0.417)	Ukraine (0.417)
KZ (0.607)	Kazakhstan (0.607)
KE (0.505)	Kenya (0.505)
PL (0.665)	Poland (0.697), Czech Republic (0.453) , Slovakia (0.212) , Slovenia (0.471)
DZ (0.592)	Algeria (0.604), Morocco (0.273)
GR (0.653)	Greece (0.653)
RO (0.691)	Romania (0.563), Albania (0.656), Hungary (0.623), Republic of Moldova (0.409)
VN (0.883)	Vietnam (0.883)
IS (0.921)	Iceland (0.921)
RS (0.822)	Serbia (0.822)
BG (0.755)	Bulgaria (0.755)
MK (0.726)	Republic of North Macedonia (0.726)
RU (0.605)	Russian Federation (0.605)
LT (0.770)	Lithuania (0.770)
FI (0.710)	Finland (0.710)
NG (0.614)	Nigeria (0.614)
LV (0.759)	Latvia (0.759)
CY (0.652)	Cyprus (0.652)
YE (0.497)	Yemen (0.497)

For measuring the performance of ethnic classification, I use test data from Behind the Name [Beh]. This site serves the unique surname and the meaning of 59 ethnic categories (they call it 'usage'). The ethnic categories are different from 'ethnic groups' in this study. To compare the classification result of my model, I use ethnic categories with the same class as the ethnic group in this study. In addition, in the case of ethnic categories that are more subdivided than my ethnic group, several ethnic categories are combined as one group. In the opposite case, ethnic groups were grouped into one ethnic category. The ethnic categories/groups used for testing are as shown in table 6.2. There are surnames that are used repeatedly by other ethnic categories. For example, surname such as Lee (English surname) and Lee (Korean surname) are excluded. A total of 4,421 surnames and ethnic pairs are used to conduct the test.

Table 6.2: Comparison of test data and ethnic groups in this study. The number of test data for each class is shown in parentheses.

Ethnic categories from test dataset	Ethnic groups
West European (971)	DE, AT
Arabic (33)	EG
Nordic (311)	SE
Hispanic (318)	CO, BR
Korean (26)	KR
African (25)	KE, NG
Japanese (193)	JP
East European (74)	PL, RU
English (1,650)	GB
Chinese (82)	CN
Indian (46)	ID
Italian (692)	IT

Table 6.3 shows the test results. The precision, recall, and F1 value for each ethnic categories are displayed. The average accuracy is 79.6%. The results in parentheses are random guesses and are displayed to compare the classification model's performance.

In the case of Korean, the degree of classification was low, unlike the previous test of the model using ORBIS dataset. The reason is the notation problem of the Behind the name dataset. I find that the notation for transferring pronunciation to the alphabet differs from the dataset used to construct a classification model. For example, the surname Mun(문) is written as both Moon and Mun in the test dataset, but in the case of Moon, it was not correctly classified. The same errors occur in the case of Kim(김) and Gim(김) Kang(강) and Gang(강). In this way, it is revealed that classification errors may occur due to differences in expressing the pronunciation of each language in alphabets.

Table 6.3: Test results of ethnicity classification

Ethnic categories	Precision	recall	F1
West European	0.77 (0.22)	0.74 (0.08)	0.76 (0.12)
Arabic	0.29 (0.01)	0.89 (0.08)	0.44 (0.02)
Nordic	0.66 (0.08)	0.86 (0.08)	0.75 (0.08)
Hispanic	0.72 (0.08)	0.79 (0.08)	0.75 (0.08)
Korean	0.83 (0.00)	0.58 (0.08)	0.68 (0.00)
African	0.7 (0.0)	0.84 (0.08)	0.76 (0.01)
Japanese	0.98 (0.04)	0.93 (0.08)	0.95 (0.05)
East European	0.37 (0.02)	0.98 (0.08)	0.54 (0.04)
English	0.90 (0.34)	0.80 (0.08)	0.84 (0.13)
Chinese	0.75 (0.02)	1 (0.08)	0.86 (0.04)
Indian	0.22 (0.03)	0.64 (0.08)	0.33 (0.04)
Italian	0.95 (0.12)	0.79 (0.08)	0.86 (0.1)
avg	0.68 (0.08)	0.82 (0.08)	0.71 (0.06)

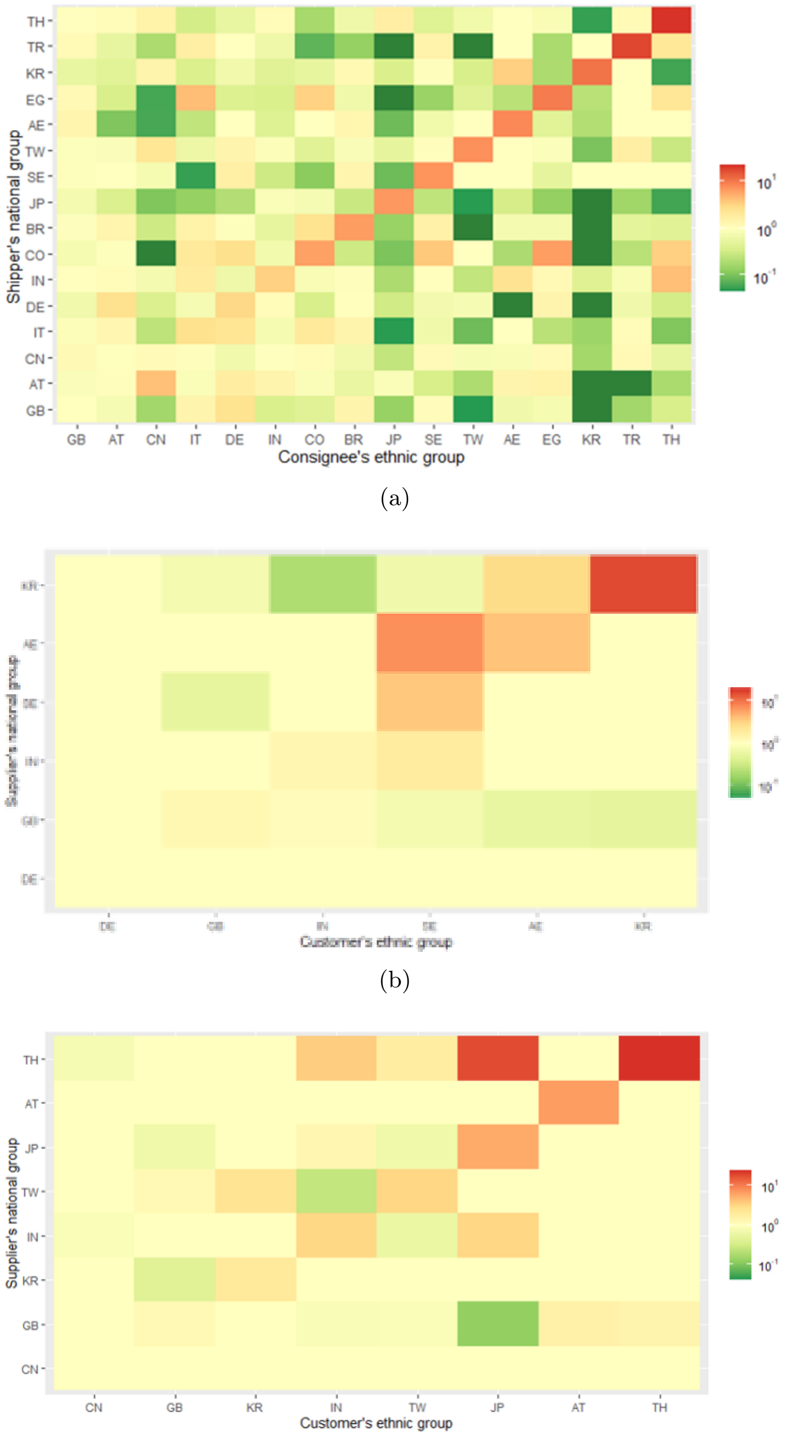
6.1.2 Ethnic homophily in international trade

I determined the majority ethnic group for each of the 662,742 companies that contain information on over three managers. For example, companies with more than a majority of managers classified as Japanese are labeled as the JP group. To investigate commerce between ethnic groups, I classified the importing and exporting companies into Top 16 ethnic groups in table 5.1. I calculated the conditional probability of trade transactions to visualize how significant transactions are made between identical ethnicities. The conditional probability of trade between ethnic groups is calculated as follows:

$$P(T) = \frac{P(A|B)}{P(A)} \tag{6.1}$$

where $P(A|B)$ is the conditional probability that group B imports from group A and $P(A)$ is the unconditional probability of importing from group A. I confirmed the independence of the unconditional probability of trading and its conditional probability by a z-test. After that, I expressed in a heatmap the value of $P(T)$, which was verified as independent ($p < 0.05$). Yellow denotes there is no significant number of transactions between the ethnic groups, and red denotes a significantly high number of transactions. A significantly low number of transactions is marked in green.

Figure6.2a shows how actively ethnic groups in the U.S. trade with the same ethnic group. Ethnic homophily is observed in most ethnic groups. The Asia JP, TW, KR, TH and Middle East AE, EG,TR ethnic groups have an especially strong tendency to trade with the same ethnic group. Such ethnic homophily in international trade is also commonly observed in interfirm transactions with non-US countries. Figures 6.2b and 6.2c represent ethnic activities in international trade with countries that belong to the DE and CN groups. For these countries, unlike the U.S., the dataset does not cover all shipping records, although I identified a trend of active transactions between identical ethnic groups.



(a)

(b)

(c)

Figure 6.2: Frequency of transactions between ethnic groups: (a) consignees in U.S., (b) companies belonging to DE group, and (c) companies belonging to CN group.

6.1.3 Measuring ethnic factors in international trade

I statistically investigated the contribution of ethnic connections to international trade using the Gravity model. The standard Gravity model for trade describes the amount of bilateral trade flow in terms of economic size and distance. I measured the contribution of ethnic factors by adding them to the Gravity model. I use the FactSet shipping and WTFC trade datasets in this section. The GDP and distance values for each country are taken from CEPII's gravity dataset. First, I used the FactSet shipping dataset, which provides the ethnic groups of each company in the U.S. By investigating the statistical significance of the ethnic factors in U.S. international trading I can identify the differences by ethnicity. Next I analyzed the statistical significance of the ethnic factors on a global scale from the WTFC trade dataset and checked whether ethnicity is statistically significant even when language barriers are removed.

6.1.4 Gravity model

The following is the standard Gravity model:

$$\ln F_{ij} = \beta_0 + \beta_1 \ln M_i + \beta_2 \ln M_j + \beta_3 lnD_{ij} + \epsilon_{ij}, \tag{6.2}$$

where i and j denote exporting and importing countries. F_{ij} is the volume of trade in US dollars between countries i and j. M_i and M_j are the GDPs in US dollars of each country. D_{ij} is the distance measured based on the most population-dense city of each country. ϵ_{ij} is an error term. We added ethnic factor E_{ij} to this standard Gravity model:

$$\ln F_{ij} = \beta_0 + \beta_1 \ln M_i + \beta_2 \ln M_j + \beta_3 lnD_{ij} + \beta_4 E_{ij} + \epsilon_{ij}, \tag{6.3}$$

where E_{ij} represents the cosine similarity in the ethnic distribution between countries i and j:

$$E_{ij} = \frac{IJ}{||I||||J||} = \frac{\sum_{k=1}^{n} I_k J_k}{\sqrt{\sum_{k=1}^{n} I_k^2}\sqrt{\sum_{k=1}^{n} J_k^2}}, \tag{6.4}$$

where vectors I and J are the ethnic composition of countries i and j. I_k is a component of vector I, and $n = 35$ is the number of ethnic groups. The ethnic distribution (vector) of corporate managers in each country is obtained from the ORBIS dataset. For example, the cosine similarity between the ethnic composition of the U.S. and the U.K. is 0.9772 and 0.0565 between the U.S. and China.

The Gravity model of trade between country i and ethnic group k within country j can be described by extending Eq. 6.3:

$$lnF_{i,k\in j} = \beta_0 + \beta_1 \ln M_i + \beta_2 \ln M_{k\in j} + \beta_3 lnD_{ij} + \beta_4 E_{i,k\in j} + \epsilon_{i,k\in j}, \tag{6.5}$$

where $F_{i,k\in j}$ is the volume of trade in US dollars between country i and ethnic group k within country j. We assume the following GDP of ethnic group k in country j: $M_{k\in j)}=$ (rate of company's majority ethnic group k in country j) / (total GDP of country j). $E_{i,k\in j)}$ represents the similarity of the ethnic composition between ethnic group k within countries j and i. The statistical significance of the ethnic factor in international trade is clarified by model selection based on Akaike's Information Criterion (AIC) and regressions on these three Gravity models, Eqs. 6.2, 6.3, and 6.5.

6.1.5 Statistical significance of ethnic factor in international trade

Using the Gravity models introduced in the previous section, I show that ethnic factors are statistically significant in international trade. In particular, I confirmed their significance for both imports and exports with Asian countries. In the multi-ethnic United States, companies in Asian ethnic groups tend to choose Asian countries as trading partners, and companies in non-Asian ethnic groups avoid choosing specific countries. Such ethnicity is even confirmed among ethnic groups that speak a common language.

I investigated whether the Gravity model with/without an ethnic factor (Eqs. 6.2 or 6.3) is chosen based on AIC for all the combinations of countries included in the WTFC trade dataset. I selected the Gravity model with an ethnic factor, which is statistically significant (Table6.4)

Table 6.4: Regression coefficients and AIC for Gravity models with/without ethnic factors, Eqs. 6.2 and 6.3, calculated for all combinations of exporting and importing countries recorded in WTFC trade dataset. (***p-value $<$ 0.001, **p-value $<$ 0.01, * p-value $<$ 0.05, p-value $<$ 0.1.)

Gravity model	Constant β_0	GDP of exporter β_1	GDP of importer β_2	Distance β_3	Ethnicity β_4	AIC
Without ethnic factor	-13.36***	0.91***	0.86***	-0.98***		93245
With ethnic factor	-13.47***	0.91***	0.85***	-0.94***	**0.51*** **	**93216**

I clarified the country dependence of the ethnic factor significance by fixing exporting country i or importing country j in the Gravity models. Here I use both the WTFC trade dataset and the FactSet shipping dataset, which shows the number of shipping containers imported by the United States. Table 6.5 shows the regression coefficients and the AIC of the Gravity models when importing country j is fixed. China and Korea tend to import from countries with a similar ethnic composition. In other countries, the Gravity model with an ethnic factor is not supported by the AIC. I next display the regression coefficients and the AIC of the Gravity models when exporting country i is fixed in Table 6.6. For imports, China and Korea export from countries with a similar ethnic composition. Japan and Spain show similar features. On the other hand, no significant ethnic effect is observed in exports from the U.S., Germany, and Brazil.

Table 6.5: Regression coefficients and AIC of Gravity models (Eqs. 6.2,6.3 and 6.5) under condition where importing country j is fixed. Coefficients were estimated using WTFC trade and FactSet shipping datasets. (***p-value < 0.001, **p-value < 0.01, * p-value < 0.05, p-value < 0.1.)

Dataset	Importer j	Gravity model	Constant $\beta_0 +$ $\beta_2 \ln M_j$	GDP of exporter β_1	Distance β_3	Ethnicity β_4	AIC
FactSet	U.S.	Without ethnic factor	-28.49***	1.32***	1.38**		1112.5
		With ethnic factor	-26.77***	1.60***	0.72	-2.29*	1122.6
WTFC	U.S.	Without ethnic factor	1.80	1.22***	-0.30		2744.6
		With ethnic factor	1.81	1.22***	-0.30	-0.003	2746.6
	Germany	Without ethnic factor	13.45***	0.77***	-0.75***		2689.7
		With ethnic factor	13.63***	0.79***	-0.81***	-0.47	2691
	Spain	Without ethnic factor	13.23***	0.78***	-0.88***		2539.7
		With ethnic factor	13.37***	0.78***	-0.90***	0.45	2539.6
	Brazil	Without ethnic factor	5.19.	1.15***	-0.80**		2438.5
		With ethnic factor	4.78	1.14***	-0.75*	0.13	2440.4
	Japan	Without ethnic factor	10.63**	1.12***	-1.22***		2587.2
		With ethnic factor	10.90**	1.11***	-1.22***	3.56	2589
	China	Without ethnic factor	5.16	1.18***	-0.65**		2691.4
		With ethnic factor	2.52	1.18***	-0.38	**1.78****	**2682.7**
	Rep. of	Without ethnic factor	8.99**	1.00***	-0.83***		2545.7
	Korea	With ethnic factor	5.01.	1.01***	-0.43	**13.27****	**2540**

Table 6.6: Regression coefficients and AIC of Gravity models, Eqs. 6.2 and 6.3, under condition where exporting country i is fixed. Coefficients were estimated using WTFC trade dataset. (***p-value $<$ 0.001, **p-value $<$ 0.01, * p-value $<$ 0.05, p-value $<$ 0.1.)

Importer i	Gravity model	Constant β_0+ $\beta_1 \ln M_j$	GDP of exporter β_2	Distance β_3	Ethnicity β_4	AIC
U.S.	Without ethnic factor	3.09	1.17***	-0.38		2710.1
	With ethnic factor	2.46	1.17***	-0.31	0.38	2711.4
Germany	Without ethnic factor	12.34***	0.81***	-0.67***		2710.2
	With ethnic factor	12.35***	0.82***	-0.69	-0.22	2711.9
Spain	Without ethnic factor	13.86***	0.78***	-0.98***		2518.6
	With ethnic factor	14.33***	0.76***	-1.01***	**0.92****	**2508.8**
Brazil	Without ethnic factor	7.49**	0.97***	-0.63*		2493.4
	With ethnic factor	6.86*	0.97***	-0.55	0.20	2495.3
Japan	Without ethnic factor	16.82***	0.99***	-1.60***		2590.3
	With ethnic factor	16.98***	0.97***	-1.63***	**13.44***	**2588.4**
China	Without ethnic factor	14.05***	0.94***	-1.04***		2763.7
	With ethnic factor	9.91***	0.93***	-0.57***	**1.87*****	**2739.6**
Rep. of Korea	Without ethnic factor	18.49***	0.77***	-1.33***		2595
	With ethnic factor	13.46***	0.74***	-0.73**	**15.20*****	**2587**

As shown in tables 6.5 and 6.6, the U.S. and Europe do not tend to choose countries with a similar ethnic composition as trading partners, reflecting a feature of their entire multi-cultural societies. However, companies in Asian ethnic groups, which are minorities in U.S. and Europe, often do choose countries of identical ethnicity as trading partners. Table 6.7 shows the regression coefficients and the AIC of the Gravity model (Eq.6.5) for each ethnic group in the United States. In ethnic groups CN, JP, TW, EG, KR, and TH, the Gravity model with ethnic factors is selected by a lower AIC score, and the factor's coefficient is statistically significant. These results are consistent with the matrix of trade frequencies among ethnic groups in figure 6.2a. The active commerce among identical ethnic groups in figure 6.2a, which cannot be explained by GDP and distance, shows the importance of the ethnic factor in trade, especially for Asian ethnic groups.

Table 6.7: Regression coefficients and AIC of Gravity model, Eq.6.5, for each ethnic group in United States. Coefficients were estimated using FactSet shipping dataset. Codes for ethnic groups are identical as in table 6.1 and Figure 6.2a (***p-value $<$ 0.001, **p-value $<$ 0.01, * p-value $<$ 0.05, p-value $<$ 0.1.)

Ethnic group k in j=U.S.	Gravity model	Constant β_0+ $\beta_2 \ln M_{k\in j}$	GDP of exporter β_1	Distance β_3	Ethnicity β_4	AIC
GB	Without ethnic factor	-0.48	1.11*	-1.31		901
	With ethnic factor	2.49	1.07*	-1.55	-0.36	905
AT	Without ethnic factor	-21.58***	1.12***	0.75		667.32
	With ethnic factor	-21.67***	1.12***	0.75	-0.11	669.32
CN	Without ethnic factor	-15.72*	0.89**	0.54		548.96
	With ethnic factor	-13.52.	1.08***	-0.23	**3.07*** **	**539.55**
IT	Without ethnic factor	-16.14*	0.69**	1.07**		591.33
	With ethnic factor	-14.11*	0.68**	0.88*	-3.30	591.54
DE	Without ethnic factor	-20.29***	1.09***	0.72*		729.06
	With ethnic factor	-20.27***	1.08***	0.73	0.01	731.06
IN	Without ethnic factor	-20.19***	1.04***	0.68		578.18
	With ethnic factor	-16.31**	1.12***	0.06	0.31	583.05
CO	Without ethnic factor	-12.00.	1.01***	-0.27		475.64

	With ethnic factor	-16.53	1.09***	0.03	1.84	475.98
BR	Without ethnic factor	-9.60	0.70**	0.29		555.64
	With ethnic factor	-11.66	0.75***	0.38	0.79	555.78
JP	Without ethnic factor	-23.30**	1.27***	0.57		573.99
	With ethnic factor	-0.86	0.70 *	-0.64	**3.43**	**572.83**
SE	Without ethnic factor	-27.06***	1.13***	1.23**		606
	With ethnic factor	-26.59***	1.12***	1.20**	-0.28	607.91
TW	Without ethnic factor	1.23	0.2278	0.2719		531.94
	With ethnic factor	-11.22*	0.94***	-0.23	**5.76*** **	**417.21**
AE	Without ethnic factor	-8.40	0.79**	-0.40		251.45
	With ethnic factor	-6.44	0.70*	-0.40	**-2.72**	**253.08**
EG	Without ethnic factor	-2.11	0.05	0.87		441.73
	With ethnic factor	-11.21	0.55	0.64	**3.80***	**429.32**
KR	Without ethnic factor	-42.84***	2.22***	0.53		395.76
	With ethnic factor	-5.55	0.85**	-0.69	**6.20*** **	**374.99**
TR	Without ethnic factor	-19.79*	0.82**	0.89*		340.23

	With ethnic factor	-20.02*	0.83**	0.90*	0.80	341.78
TH	Without ethnic factor	1.94	0.42	-0.55		377.71
	With ethnic factor	-10.80.	0.89***	-0.33	**4.49*** **	**352**

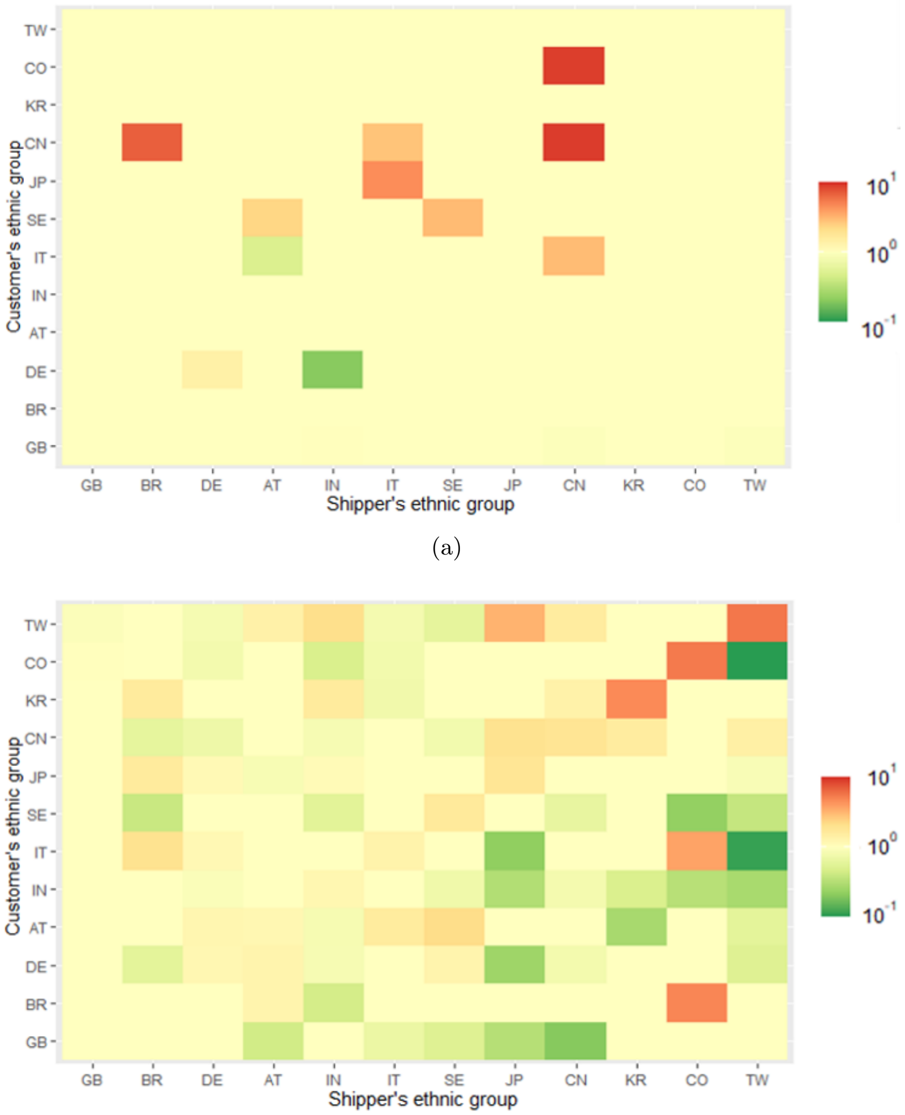
The same ethnic group often speaks the same language. I show that the ethnic factor contains more than just language barriers for trading. To evaluate the ethnic factor's significance, I focused on commerce between ethnic groups and countries where more than 9% of the population speaks the same language (It includes countries that use the same primary language and countries with a colonial relationship). Here I assume that the people of ethnic group k in country j can understand the primary language of its country and the language of its ethnic group. Table 6.8 shows the regression coefficients and the AIC of the Gravity models (Eqs.6.2, 6.3, and 6.5) for commerce between such countries and ethnic groups that understand the same language. Even if we remove the language barrier, the ethnic factor is significant in international trade and shipping.

Table 6.8: Regression coefficients and AIC of Gravity models (Eqs.6.2, 6.3, and 6.5) for trade between such countries and ethnic groups that understand same language. Coefficients were estimated using WTFC trade and FactSet shipping datasets. (***p-value $<$ 0.001, **p-value $<$ 0.01, * p-value $<$ 0.05, p-value $<$ 0.1.)

	Model	Constant β_0	GDP of exporter β_1	GDP of importer β_2	Distance β_3	Ethnicity β_4	AIC
Trade	Without ethnic factor	-7.95***	0.80***	0.75***	-1.04***		12671
	With ethnic factor	-8.24***	0.79***	0.73***	-0.97***	**0.46***	**12666**
Shipping to ethnic group k, in U.S.	Without ethnic factor	-38.98***	1.42***	0.67***	0.49***		5218.5
	With ethnic factor	-39.38***	1.37***	0.69***	0.57***	**0.53***	**5216.4**

6.2 The difference between domestic trade and international trade

The tendency to prefer the same ethnicity is a feature unique of international trade that is not observed in domestic trade. FactSet's Supply Chain dataset contains domestic and international trade information between companies. Figure6.3 shows the frequency of transactions between ethnic groups in domestic and international trade by estimating the conditional probability (Eq.6.2) from this dataset. Here I removed the language barrier by limiting the transactions to the following English-speaking countries: Australia, Canada, Ireland, New Zealand, South Africa, United Kingdom, and United States. In domestic transactions, I notice no homophily feature (where companies prefer to choose companies of the same ethnicity as their trading partners) is present except in the CN, SE ethnic group. On the other hand, the reddish color around the 45-degree line on the heatmap (Figure6.3b) indicates active connections between companies in the same ethnic group in international trade. The significance of the ethnic factor in trading, which has been clarified up until this section, is a characteristic unique to international trade.



(a)

(b)

Figure 6.3: Frequency of transactions between ethnic groups in domestic and international trade in English-speaking countries, Australia, Canada, Ireland, New Zealand, South Africa, United Kingdom, and United States: (a) and (b) domestic and international trades

6.3 Case study

The development of technology and transportation has fueled much globalization in modern societies. Such globalization refers mainly to the interaction and economic integration process related to social and cultural aspects. But our study is based on the general idea that the simple movement of people may cause globalization. In other words, rather than interaction with the world's society, culture, and economy, I assume that the illusion of globalization is caused by the people who emigrated and settled in countries around the world. To verify this assumption, I first observed international trade exchanges through U.S. trade data with sufficient statistical data. I compared general international trade data in circumstances in which trade impediments occurred and analyzed how ethnic networks affected these situations. If an ethnic network is activated during a trade failure, it suggests the significance of its impact on the substance in international trade activities. I analyzed the data from the recent Arab Spring and the U.S.-China trade war as obstacles to trade with the U.S.

U.S.-China trade war analysis

I need the data which can address the international trading aspects of the United States. FactSet is appropriate dataset to show the flow of international trading. The FactSet data include trade data from each country that interacts with the United States. Through a combination with the business people data that have already been secured, we can analyze how ethnic networks affect trade between countries, especially during the recent trade war between the United States and China. The U.S. international trading data provided by FactSet include the following dataset. The trading (import) record identifies transactions from Jan. 2014 to Mar. 2019. A dataset of FactSet entities, FactSet Person entities, was prepared for analysis. FactSet Person entities provide personal names and the companies at which they are employed. I used the names of company executives for classifying ethnic groups with a surname-ethnicity classifier for identifying ethnic networks in U.S. international trading.

The FactSet trading record dataset was cross-checked with U.S. Census trading records. The U.S. Census provides U.S. international trading data with port, commodity, country, and time stamps. Since it only shows the trading value, the FactSet trading record was modified into value unit records.

However, the FactSet trading dataset includes 41% null value records. Instead of value data, I used the number of records and estimated the price index with the Dow Jones Iron Steel Index (DJUSST), which includes the stock prices of major steel and metal-related companies. It is used by asset management companies for trading metal-related assets. Since the trade record of metal is stable in all trading history, it can be used as an index of trading datasets. The U.S. international trading records were multiplied by this index. The cross-check result between the U.S. Census and the FactSet dataset is shown in figure 6.4.

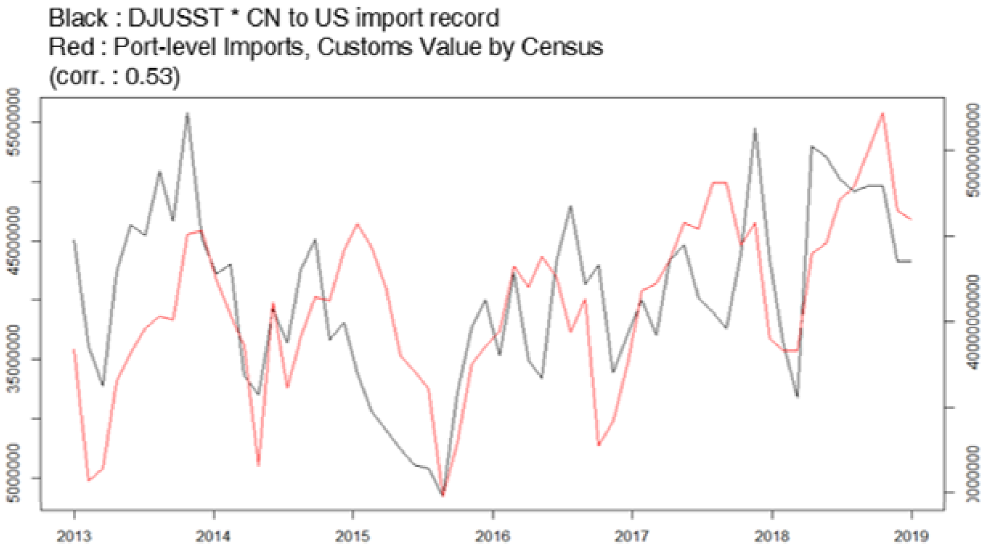


Figure 6.4: Relationship between Census and FactSet data

I analyzed the influence of ethnic networks observed in the 2018 trade war between the U.S. and China that began in August 2017 when the United States Trade Representative (USTR) start the investigation China's IP and tech transfer policies under Section 301 tariffs. Section 301 of the Trade Act of 1974 grants the Office of the USTR a range of responsibilities and authorities to investigate and take action to enforce U.S. rights under trade agreements and respond to certain foreign trade practices[cit]. Section 301 provides a statutory means by which the United Sates. imposes trade sanctions on foreign countries that violate U.S. trade agreements or engage in acts that are "unjustifiable" or "unreasonable" and burden U.S. commerce. Prior to 1995, the U.S. used Section 301 extensivelyto pressure other countires to eliminate trade barriers and open their markets to

U.S. exports[cit]. Since Section 301 enactment in 1974, there have been 130 cases under the law, of which 35 have beem initiated since the WTO's establishement in 1995. These cases have primarily targeted the European Union, Canada, Japan, and South Korea. Prior to 2017, the last Section 301 investigation took place in 2013 and involved Ucraine's practices regarding IPR. The last investigation prior to the Trump Administration resulting in retaliation (i.e., tariffs) took place in 2009 and involved Canada's compliance with the 2006 U.S. - Canada softwood Lumber Agreement. During the Trump Administration, the USTR initiated 6 new investigations: China (Aug. 2017), EU(April 2019), France(July 2019), Foreign Digital Services Taxes(July 2020), Vietnam(Oct. 2020)[cit]. In this paper, I analyse China's case because its period is included in my dataset.

Figure 6.5 shows the timeline of Section 301 in Trump Administration. In August 2017, Trump orders Section 301 probe into alleged Chinese intellectual property theft. USTR start to investigate China's IP and tech transfer policies. Trump announces Section 301 List 1 and 2 in March 2018, and propose expanding Section 301 in April 2018. From July 2018, the tariffs start to go into effect. Table 6.9 shows tariffs along to each List.

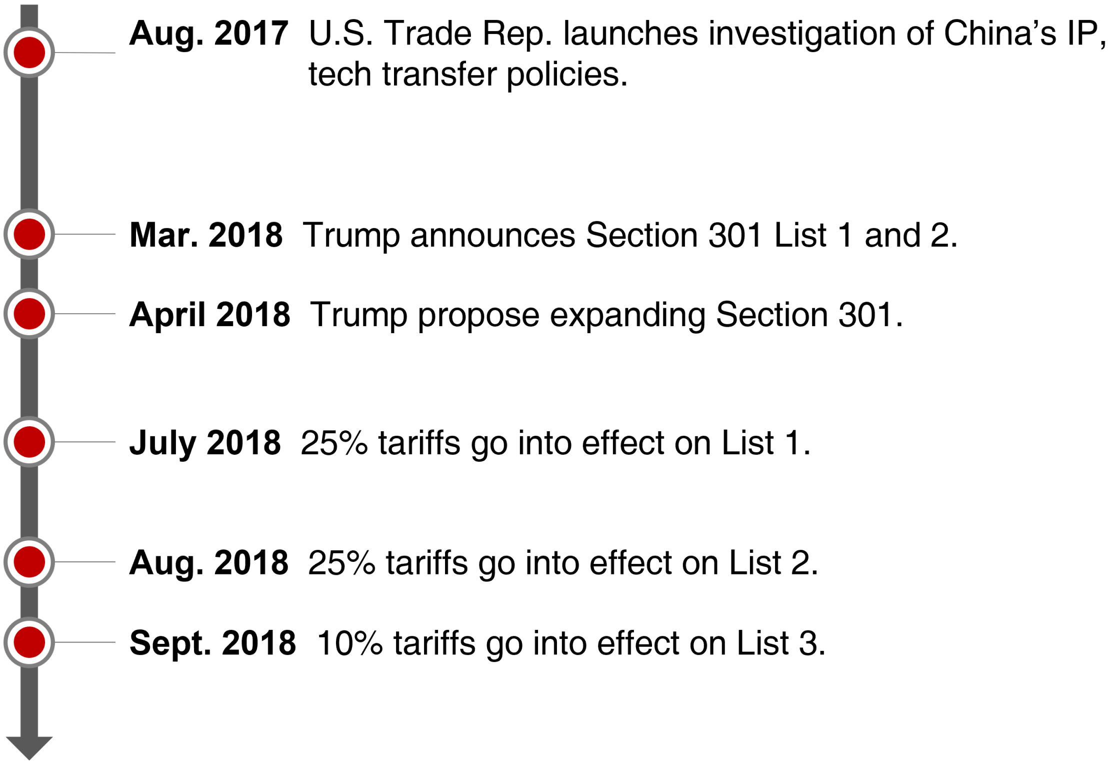

Aug. 2017 U.S. Trade Rep. launches investigation of China's IP, tech transfer policies.
Mar. 2018 Trump announces Section 301 List 1 and 2.
April 2018 Trump propose expanding Section 301.
July 2018 25% tariffs go into effect on List 1.
Aug. 2018 25% tariffs go into effect on List 2.
Sept. 2018 10% tariffs go into effect on List 3.

Figure 6.5: Timeline on Section 301 tariffs on Chinese goods from 2017 to 2018.

Table 6.9: Section 301 tariffs on each list [lis]

List 1 - $34 Billion U.S. Goods / 545 Products
Effective July 6, 2018 25% Additional Duty Automobiles Soybeans, Corn, Wheat, Rice Sorghum, Tobacco, Cigars, Alcohol, Dog/Cat Food Beef, Port, Poultry, Fish/Shellfish, Dairy products, Nuts, Vegetables
List 2 - $16 Billion U.S. Goods / 333 Products
Effective : August 23, 2018 25% Additional Duty Cole, Coke(fuel), Crude Oil, Diesel and Kerosene Gas Textile materials, Metals Waste/Scrap, Other Vehicles, Motorcycles Medical Equipment
List 3 - $60 Billion U.S. Goods / 5,207 Products
Effective : September 24, 2018 Additional Duties : Annex 1 - 10% increased to 25% Annex 2 - 5% incrased to 20% Annex 3 - 5% increased to 10% Annex 4 - 5% with no increase Honey, Veg/Plant Oils, Sugars, Coffee and Tea, Spices, Plastic Leather, wood and paper products Home and industrial machines and Electronics
List 4 - $75 Billion U.S. Goods / 5,079 Products
Effective : September 1,2019 10% on 916 products in Annex 1 5% on 801 products in Annex 1 Effective December 15, 2019 10% on 913 products in Annex 2 5% on 2,449 products in Annex 2 Effective February 14, 2020 Tariffs reduced by 50% on 1,717 products

From 2014 to 2019, total imports from China to the U.S. are relatively stable (Figure 6.6). I investigated how the impact of the Chinese ethnic network is represented when Section 301 tariffs act with data from company executives whose ethnic origin was previously classified. I only selected the data of U.S. consignees that have executives who were classified as Chinese to observe the trading records of the same period.

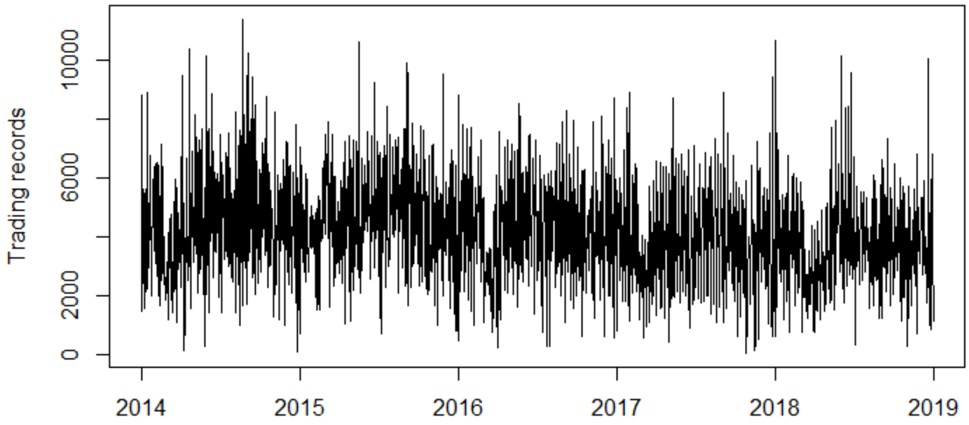


Figure 6.6: Trade record of U.S. imports from China The x-axis and the y-axis indicate the day and the trading records, respectively

To identify how the trade of a company with a Chinese executive changed over the total trading volume during the period, I divided the number of trading records by the total trade amount. Figure 6.7 shows the ratio of the trade volume of companies with Chinese executives in the overall commodity sector. It was obtained by dividing the trade records of U.S. companies in which Chinese executives were employed by the total trade records with China (excluding NA values). The left y-axis represents the rate, and the right y-axis represents the cumulative rate. I executed a two sample t-test to check the differences 6 month before and after the trade war(announcement of the Section301, before: 2016-02-14 2017-08-13, after:2017-08-14 2018-02-13). I observed significantly different between the two periods and the difference (after - before) is greater than 0. (p-value <0.01).

I have identified visible fluctuations, especially in the retail sector. Figure 6.8 shows the ratio of trade volume in the total retail sector of companies with a Chinese executive among consignees in the retail sector. Significant differences were also observed in the t-test of the retail sector, 6 month before and after the announcement of the Section301 (before: 2016-02-14 2017-08-13, after:2017-08-14 2018-02-13). I can't observe any significant difference between two period. But with 1 year period(before: 2016-08-14 2017-08-13, after:2017-08-14 2018-08-13), there's significant difference between two period, and the difference is greater than 0 (p-value<0.01). For reference, the American government also imposed tariffs on China in 2014. The role of ethnic networks is activated when obstacles arise in trade between countries.

In the retail sector, the impact of ethnic networks is slightly greater than in other areas. For areas where reference prices have not been formed in trade transactions, a related study [Tri02] found that the influence of ethnic networks can be greatly affected. In areas with no existing reference prices, it is critical to go beyond information barriers to enter the market. In information exchanges formed at this time, a network comprised of ethnic, linguistic, and cultural affairs is bound to play a huge role.

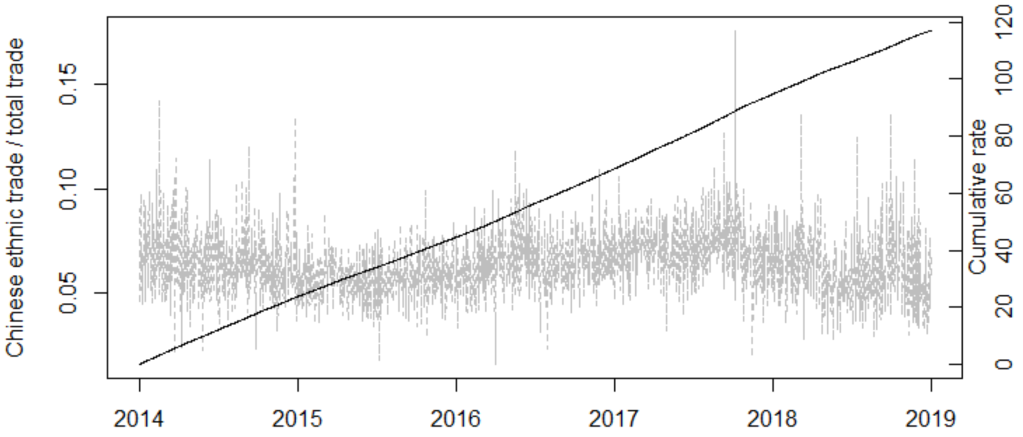


Figure 6.7: Trade record of U.S. company imports from China that employs Chinese executives (all commodities) The gray dotted line represents the rate of Chinese ethnic trade in total CN-US trade records. The black line represents the cumulative rate.

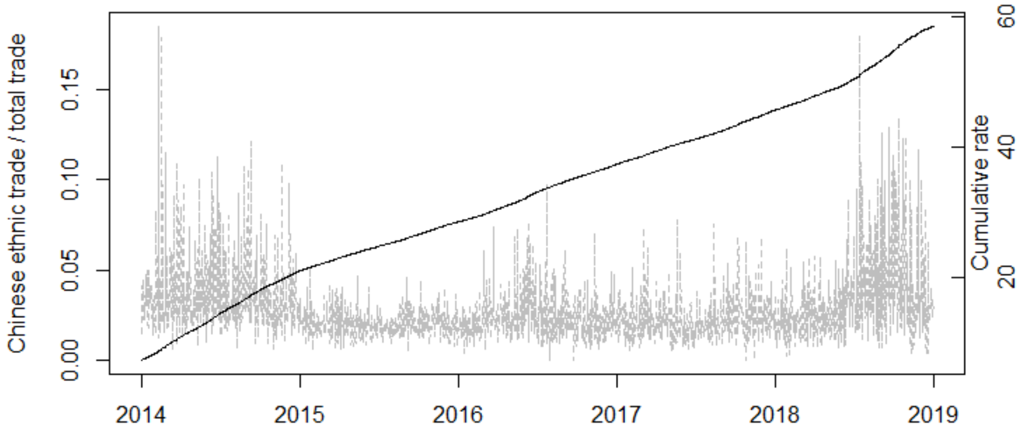


Figure 6.8: Trade record of U.S. company import from China that employs Chinese executives (related to retail) The gray dotted line represents the trading records. The black line represents the cumulative ratio of the trading records in whole trading records.

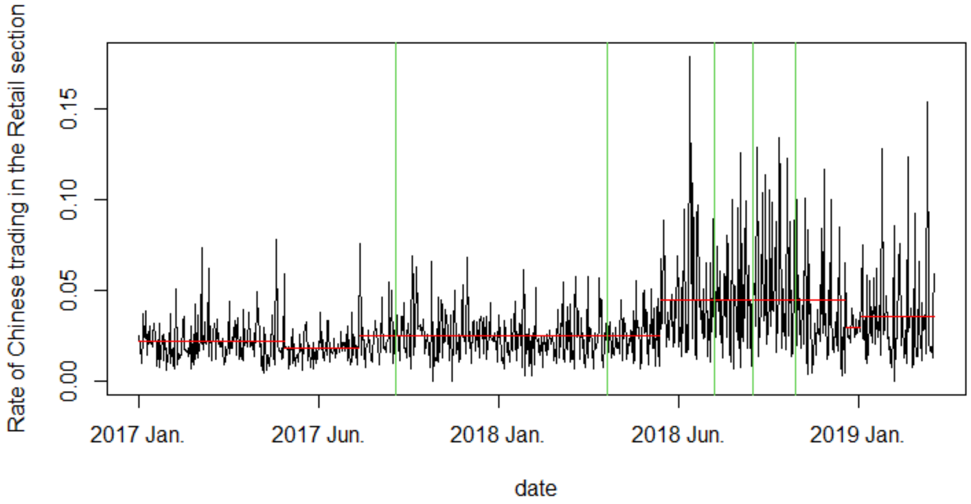


Figure 6.9: The change point of trade record: U.S. company import from China that employs Chinese executives (related to retail) The red line represents detected the change point by mean and variance. The green line represents the announcement date of Section 301.

In the retail field, I check the relationship with Section 301's timeline by focusing on the period from 2017 to 2019. For detecting change points, I use changepoint library in R. I use the cpt.meanvar function to identify the change point. This function is used to find changes in mean and variance for data using the test statistic specified in the data. The red line in figure 6.9 represents the change point of the trade record among Chinese in the retail sector. The green line is when the major announcement of section 301 tariffs was made. Considering that there is a one-month time difference in trade through ships, it can be seen that trade between the Chinese has increased since Trump proposed to expand Section 301 and began to affect lists 1-3. In other words, it can be interpreted that during this period, regulations caused problems in trade between China and the United States. To avoid this, the network between Chinese people became active.

Arab Spring

In the same way, the IHS data, which contain the U.S. international trading data during the Arab Spring, shows the influence of ethnic networks during events that affected trade between countries. The PIERS database, provided by IHS, is a database of the manually recorded bill of lading data that records the addresses of the ports loaded with cargo, country, shipper information, destination information, consignee company information, commodity details, price, etc. We used the trading data from Arab countries and the U.S. to analyze how ethnic networks were activated during the Arab Spring (2010 to 2013).

The Arab Spring was an anti-government movement that took place over two years from December 2010 to December 2012. During this period, Muslim countries in Arab and North Africa experience social chaos. Although slightly different from a trade war, in the presence of social turmoil, trade deals are affected. We analyzed this period's data on trade deals with the U.S. by specific countries and identified how ethnicity networks are activated when obstacles interrupt trade deals. Among the countries that experienced social turmoil during the Arab Spring, we analyzed Saudi Arabia and Jordan. Saudi Arabia is the U.S's 20^{th} largest trading partner, exporting mostly mineral fuels and metal to the U.S. The complete trading records between Saudi Arabia and the United States from 2010 to 2013 is illustrated in Figure 6.10. I analyzed how Arab ethnic networks were activated during and after the Arab Spring. Consignees, which include officials classified as Arabian (Algeria, Egypt, Saudi Arabia, and the United Arab Emirates) by name-ethnic classifiers were used for analyzing the effect of ethnic networks. Figure 6.11 shows how much total trade the Arab ethnic networks accounted for during the Arab Spring. I checked the differences with the trade records after, based on 2010 data before the Arab Spring. 2011 immediately after the Arab Spring was observed to have significant differences from 2010's data (p-value<0.01). It can be interpreted as evidence of the activity of the Arab ethnic linkage during the Arab Spring. Figure 6.12 also shows the change point in the trade records of Saudi Arabia and the United States. The red line is the change point of the trade record, and the green line marks the date of the big protests in Saudi Arabia. In the Arab Spring example, it can also be seen that trade between the Arabian becomes active when international trade is disrupted.

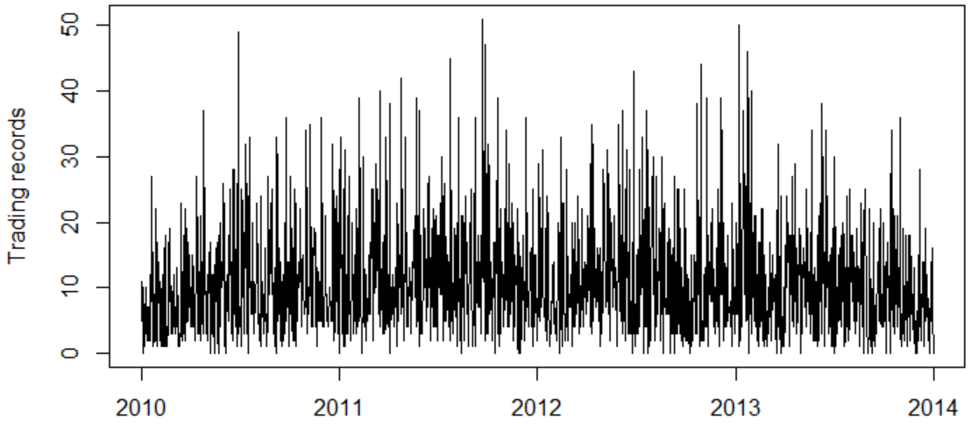


Figure 6.10: U.S. imports from Saudi Arabia The x-axis and the y-axis indicate the day and the trading records, respectively.

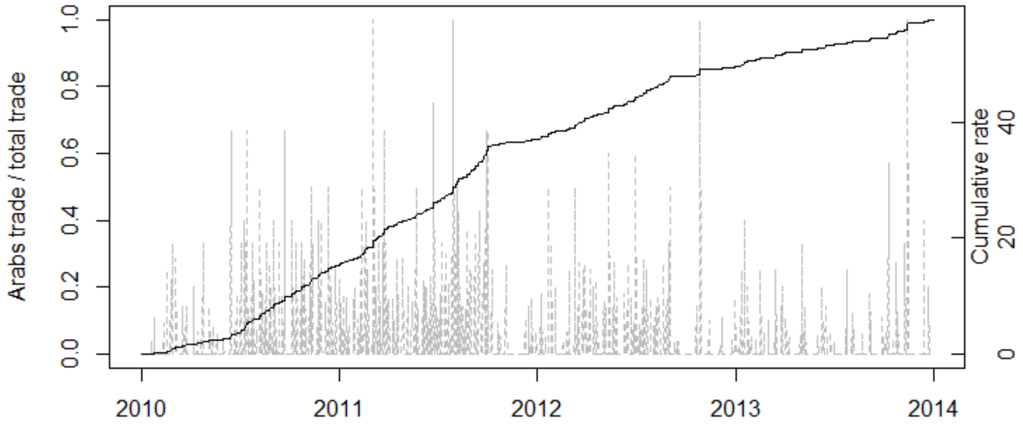


Figure 6.11: U.S. company imports from Saudi Arabia that employs Arabian executives.The gray dotted line represents the trading records. The black line represents the cumulative ratio of the trading records in whole trading records.

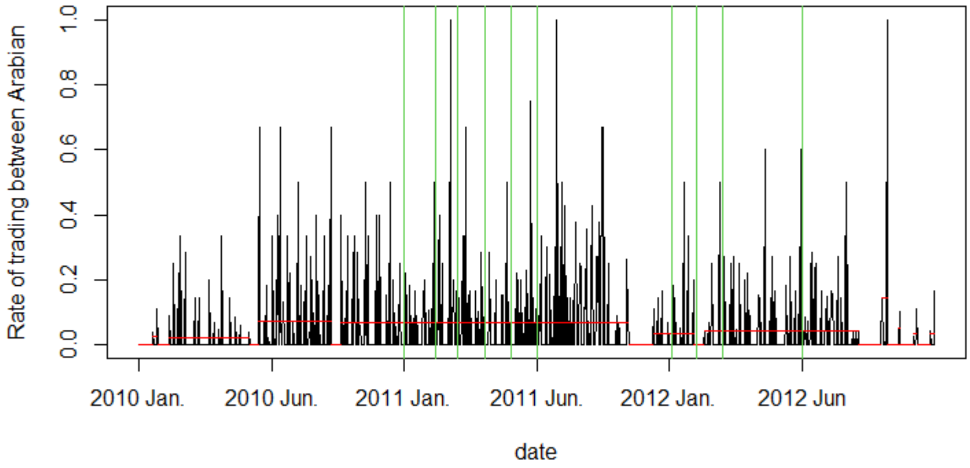


Figure 6.12: U.S. company imports from Saudi Arabia that employs Arabian executives.The gray dotted line represents the trading records. The red line represents the detected change point of the trading records in whole trading records. The green line represents the date of the big protests in Saudi Arabia.

6.4 Summary

This chapter analyzed interethnic trade in the economic realm. To this end, first of all, the origins identified through the surname-origin classifier were transformed into ethnic groups. By grouping the previously classified 77 origins into 35 ethnic groups with similar linguistic (surname) composition, classification errors were reduced, and the classified results can be interpreted as ethnicity. Using the FactSet shipping dataset, I observed distinct ethnic homophily in US international trade, especially among Asian ethnic groups. By analyzing the FactSet Supply Chain dataset in the same way, it was confirmed that interethnic trade appears in another country. I proved through the Gravity model that ethnicity is essential in international trade. Such interethnic trade is identified as a feature of international trade that is not observed in domestic trade. The last section of the chapter introduced case studies in which ethnic networks were activated in situations where trade obstacles, such as the US-China trade war and the Arab Spring, occurred.

Chapter 7

Ethical risks to consider

This study aims to understand social phenomena through ethnic classification using surnames. However, the classification of ethnic groups can raise sensitive ethical issues. This chapter explains the ethical problems of ethnic classification research, the responses in this study, and the author's opinion.

1. Re-identify unwanted individuals. The individual is an independent and perfect being in himself. Giving ethnic attributes to an individual may provide an unwanted identity to the individual. I respect that an individual has an ideal identity as it is. Ethnic classification in this study is not intended to label and discriminate against individuals but is used solely as one of the various factors to understand social members.

2. Problems caused by misclassification. Other than nationality, which can be assigned by individual choice, it is challenging to classify race, origin, and ethnicity accurately. In particular, the degree of detail of classification by ethnic group varies greatly depending on the characteristics of the data used to construct the classifier. Therefore, in this study, the classification accuracy by category of the surname-origin classifier in Chapter 3 and the classification accuracy by category of the surname-ethnic group classifier in Chapter 5 are presented. Since there are categories with low classification accuracy depending on origin or ethnic group, it is necessary not to overinterpret by referring to these categories.

3. Discrimination due to misinterpretation or making policy. The author cautions against misinterpretation of the characteristics of each ethnicity, discrimination, and wrong policy establishment using the method of ethnic classification presented in this study. Ethnic classification studies should not be used to harm anyone or to penalize a particular group. The author also conducted research with a clear understanding of this problem and was careful not to overinterpret or misinterpret the research results. I hope you will pay attention to these points when establishing policies based on this study or researching ethnic and social phenomena.

Chapter 8

Conclusion

This study developed a classifier that could verify the origin of people using large-scale surname data and the RNN method. Using this classifier, I analyzed the role of ethnicity in sociology and economics, especially international trade. First, as training data for RNN, I used the surname and nationality data of executives in economics fields obtained from the ORBIS dataset. This dataset contained large-scale data of about 35 million entities. This study is the first attempt in this field to construct a classifier using massive amounts of reliable name data. Therefore, I could obtain a classifier with high predictive precision.

I built a surname origin classifier using the training data of 77 countries having sufficient information. I then compared the predictions of this classifier with the findings of relevant studies using the surname nationality data of Olympic athletes. I also compared the prediction with the MTurk dataset, to confirm that the classification was closely related to that at the human level. Thus, the classifier could categorize ethnicity by surname at the human level.

Through an applied analysis in sociology with the built surname origin classifier, I first analyzed the ethnic diversity of each country using only surname data. This approach was meaningful in that it could overcome the problem of demographic analysis for each country without a unified standard and enabled a comparative analysis by country. I then studied the social groups in the United States and showed how the racial composition in a country differed by social group. Finally, I analyzed the spatial distribution of ethnic linkages. I then built a surname origin classifier for 48 African countries and calculated the similarities using the JSD by the origin of ethnicity distribution in each country and communities in the countries. This analysis enabled visualization of the spatial influence of ethnic linkages in the economic network. I then analyzed the spatial distribution of ethnicity in French and Swiss cities in a similar manner.

The networks of ethnicities can be observed more dynamically in economics. I investigated the ethnic homophily in international trade using corporate managers' surnames. Using the surname origin classifier for managers of these companies, I estimated the ethnicity in the companies. Thus, I could identify the composition of ethnic groups at the individual company and country levels. Evidently, the strength of homogeneity depended on ethnicity.

First, I analyzed the frequency of transactions between ethnic groups focusing on companies in the United States. Considering the majority ethnic group of each company and using the conditional probabilities to visualize the frequency of transactions between ethnic groups, I found strong ethnic homophily, especially in the Asian and Middle East ethnic groups. Next, using the gravity model to measure the statistical significance of ethnic factors in international trade, I analyzed the global trade using the WTFC dataset, to find a positive ethnic factor effect, although this differed by country. Although there was no significant ethnic factor effect in the English (GB) and European (DE) groups, positive effects could be observed in the Chinese (CN) and Korean (KR) groups. I also found significant ethnic factor effects even after removing the language barriers. A preference could be found between identical ethnic groups not in domestic, but in international trade.

I analyzed the impact of ethnic linkage in international trading as an extended application. For this, I used the US trading data obtained from FactSet and the PIERS bill of lading data obtained from IHS, and analyzed how ethnic networks in international trade are activated around specific events such as the U.S.–China trade war and the Arab Spring. This study assumed that ethnic linkage can be used to overcome the existing difficulties in international trade. I could thus conclude that exports from China to the United States decreased slightly compared to the average during the 2018 US–China trade war. However, when a Chinese executive was employed in the company of the consignee during the same period, the company's imports from China increased compared to the average. Thus, ethnic networks can be activated to overcome inter-country trade disturbances. Similarly, an analysis of some Arab countries to identify the changes in volume of transactions with the United States during the Arab Spring confirmed that in Saudi Arabia, which exports many minerals, ethnic networks in trade were active around the 2011 anti-government demonstrations.

Despite globalization, homogeneity in trade exists between several ethnic groups. Some ethnic groups trade with diverse ethnic groups. Although limiting trading partners to the same ethnic group might increase transaction costs, this might be effective against factors that impede trade, such as international conflicts and unreliable contracts.

This study has presented the results of applied research in sociology and economics done using the approach as described in the study. The method can be extended to various other fields, for example, to resolve ethnic conflicts and health care issues by ethnic characteristics. It can also be applied to the cryptocurrency market, which is expected to be active in interethnic transactions because language dependence becomes stronger to obtain information. Thus, one can expect to observe a more dynamic ethnic network in the cryptocurrency market . This method can also help in predicting the flow of cryptocurrency or tracking abnormal transactions. I leave these tasks to future works.

Bibliography

[AA09] Jahangir Mohammed Swapna Male Steven Skiena Anurag Ambekar, Charles Ward. Name-ethnicity classification from open sources. *ACM SIGKDD*, pages 49–58, 2009.

[AH19] J. W. Dembosky J. L. Adams S. M. Wilson-Frederick J. S. Mallett A. M. Haviland A. Haas, M. N. Elliott. Imputation of race/ethnicity to enable measurement of hedis performance by race/ethnicity. *Health services research*, 54:13–23, 2019.

[Ahm10] Ali M. Ahmed. What is in a surname? the role of ethnicity in economic decision making. *Applied Economics*, 42:2715–2723, 2010.

[AJC07] Richard P. Gallagher Andrew J. Coldman, Terry Braun. The classification of ethnic status using name information. *Journal of Epidemiology and Community Health*, 42:390–395, 2007.

[AM11] Y. Y. Ahn J. P. Onnela J. Rosenquist A. Mislove, S. Lehmann. Understanding the demographics of twitter users. *Proceedings of the International AAAI Conference on Web and Social Media*, 5, 2011.

[AM18] Camille Roth Antoine Mazières. Large-scale diversity estimation through surname origin inference. *Bulletin of Sociological Methodology*, 139:59–73, 2018.

[Ama] Amazon. Mturk. https://www.mturk.com. Accessed: 2023-01-02.

[AMES10] Sandro Galea Abdulrahman M. El-Sayed, Diane S. Lauderdale. Validation of an arab name algorithm in the determination of arab ancestry for use in health research. *Ethnicity health*, 15:639–647, 2010.

[AN86] S. J. Ulijaszek A. Nicoll, K. Bassett. What's in a name? accuracy of using surnames and forenames in ascribing asian ethnic identity in english populations. *JOURNAL OF EPIDEMIOLOGY AND COMMUNITY HEALTH*, 40, 1986.

[App01] Osei Appiah. Ethnic identification on adolescents' evaluations of advertisements. *Journal of Advertising Research*, 41:7–22, 2001.

[AV01] G. Biondi A. Vienna. Culture and biology: surnames in evaluating genetic relationships among the ethnic minorities of southern italy and sicily. *Collegium antropologicum*, 25:189–193, 2001.

[BB13] Sunny Lie Benjamin Bailey. The politics of names among chinese indonesians in java. *JOURNAL OF LINGUISTIC ANTHROPOLOGY*, 23, 2013.

[Beh] BehindTheName. Behindthename. https://surnames.behindthename.com/names/list. Accessed: 2023-01-02.

[BM09] Marianne Bertrand Saugato Datta Banerjee, Abhijit and Sendhil Mullainathan. Renouncing personal names: An empirical examination of surname change and earnings. *JOURNAL OF LABOR ECONOMICS*, 27:127–147, 2009.

[Bon] Bonica. http://www.stanford.edu/~bonica/. Accessed: 2023-01-02.

[BRS10] S. Amin M. Ramani S. Sadry J. V. Tu B. R. Shah, M. Chiu. Surname lists to identify south asian and chinese ethnicity from secondary data in ontario, canada: a validation study. *BMC medical research methodology*, 10:1–8, 2010.

[CEP] CEPII. Cepii dataset. http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele.asp. Accessed: 2023-01-02.

[cit] Article of section 301. https://crsreports.congress.gov/product/pdf/IF/IF11346. Accessed: 2023-01-02.

[Col05] Christian Collet. Bloc voting, polarization, and the panethnic hypothesis: The case of little saigon. *The Journal of Politics*, 67:907–933, 2005.

[Cop18] Conrad Copeland. *Bridging New Divides: The Impact of International Ethnic Linkages on Bilateral Trade in Africa*. 2018.

[Cry15] Adam Crymble. A comparative approach to identifying the irish in long eighteenth-century london. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 48:141–152, 2015.

[Daf19] Xu Dafeng. Surname-based ethnicity and ethnic segregation in the early twentieth century us. *Regional Science and Urban Economics*, 77:1–19, 2019.

[DD11] G. J. Norman V. L. Irvin D. Chhay M. F. Hovell D. Ding, C. R. Hofstetter. Measuring immigration stress of first-generation female korean immigrants in california: psychometric evaluation of demand of immigration scale. *Ethnicity Health*, 16:11–24, 2011.

[Dow] DowJones. Dow jones site. https://developer.dowjones.com/site/docs/risk_and_compliance_feeds/watchlist_ame_soc/dow_jones_watchlist/index.gsp. Accessed: 2023-01-02.

[DS10] F. Rezvani A. O. Coates D. S.Tseng, J. Kwong. Angiotensin-converting enzyme-related cough among chinese-americans. *The American journal of medicine*, 15:183, 2010.

[DSL97] S. E. Furner P. S. Levy J. A. Brody J. Goldberg D. S. Lauderdale, S. J. Jacobsen. Hip fracture incidence among elderly asian-american populations. *American journal of epidemiology*, 146:502–509, 1997.

[DSL00] Bert Kestenbaum Diane S. Lauderdale. Asian american ethnic identification by surname. *Population Research and Policy Review*, 19:283–300, 2000.

[DSL02] Bert Kestenbaum Diane S. Lauderdale. Mortality rates of elderly asian american populations based on medicare and social security data. *Demography*, 39:529–540, 2002.

[DSL07] Bert Kestenbaum Diane S. Lauderdale. Asian american ethnic identification by surname. *Population Research and Policy Review*, 19:283–300, 2007.

[DSLG96] Sylvia E. Furner Paul S. Levy Jacob A. Brody Diane S. Lauderdale, Steven J. Jacobsen and Jack Goldberg. Hip fracture incidence among elderly asian-american populations. *American journal of epidemiology*, 146, 1996.

[DW02] John Ries Don Wagner, Keith Head. Immigration and the trade of provinces. *Scottish Journal of Political Economy*, 49:507–525, 2002.

[ECW10] Diane S. Lauderdale Eric C. Wong, Latha P. Palaniappan. Using name lists to infer asian racial/ethnic subgroups in the healthcare setting. *Medical care*, pages 540–546, 2010.

[EGB03] Eliseo J Pérez Stable Dean Sheppard Esteban González Burchard, Elad Ziv. The importance of race and ethnic background in biomedical research and clinical practice. *The New England Journal of Medicine*, 348:1170–1175, 2003.

[EL06] Leeat Yariv Einav Liran. What's in a surname? the effects of surname initials on academic success. *JOURNAL OF ECONOMIC PERSPECTIVES*, 20:175–187, 2006.

[Faca] FactSet. Factset. https://www.factset.com/. Accessed: 2023-01-02.

[Facb] FactSet. Factset site. https://www.factset.com/services/data-delivery. Accessed: 2023-01-02.

[fb] fb. Cia factbook. https://www.cia.gov/the-world-factbook/countries/. Accessed: 2023-01-02.

[FL11] Pablo Mateos F. Lakha, Dermot R. Gorman. Name analysis to classify populations by ethnicity in public health: validation of onomap in scotland. *Public health*, 125:688–696, 2011.

[GJF10] Farid Toubal Gabriel J. Felbermayr, Benjamin Jung. Ethnic networks, information, and international trade: Revisiting the evidence. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, pages 41–70, 2010.

[GJM07] B. Palmer C. Afzal A. J. Silman A. Esmail G. J. Macfarlane, M. Lunt. Determining aspects of ethnicity amongst persons of south asian origin: the use of a surname-classification programme (nam pehchan). *Public health*, 121:231–236, 2007.

[Gor99] Bridget K Gorman. Racial and ethnic variation in low birthweight in the united states: individual and contextual determinants. *Health place*, 5:195–207, 1999.

[Gri18] Randi H Griffin. 120 years of olympic history: athletes and results. https://www.kaggle.com/datasets/heesoo37/

120-years-of-olympic-history-athletes-and-results, 2018. Accessed: 2023-01-02.

[HE14] Emmanuel Rocher Raju Jan Singh Hélène Ehrhart, Maëlan Le Goff. Does migration foster exports? evidence from africa. *World Bank*, 2014.

[HQ06] D. Schopflocher C. Norris P. D. Galbraith P. Faris W. A. Ghali H. Quan, F. Wang. Development and validation of a surname list to define chinese ethnicity. *Medical care*, pages 328–333, 2006.

[IB96] M. Beretta C. Nesti E. Mamolini A. Rodriguez-Larralde I. Barrai, C. Scapoli. Isonymy and the genetic structure of switzerland .1. the distributions of surnames. *ANNALS OF HUMAN BIOLOGY*, 23, 1996.

[IIW06] D. A. John R. O. Morgan I. I. Wei, B. A. Virnig. Using a spanish surname match to improve identification of hispanic women in medicare administrative data. *Health Services Research*, 41:1469–1481, 2006.

[IK16] Kabir Khanna Imai Kosuke. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, pages 263–272, 2016.

[IMZ20] Robert Rudolf Inmaculada Martinez-Zarzoso. The trade facilitation impact of the chinese diaspora. *The World Economy*, 2020.

[Iwa14] Mathias Iwanowsky. *The Role of Ethnic Networks in Africa: Evidence from Cross-Country Trade*. 2014.

[JB19] Gwilym Pryce Jessie Bakens. Homophily horizons and ethnic mover flows among homeowners in scotland. *Housing Studies*, 34:925–945, 2019.

[JC10] Lars Backstrom Cameron Marlow Jonathan Chang, Itamar Rosenn. Ethnicity on social networks. *ICWSM*, 10:18–25, 2010.

[JCA14] S. A. O'Connell M. Yang J. C. Aker, M. W. Klein. Borders, ethnicity and trade. *Journal of Development Economics*, 107:1–16, 2014.

[JCS02] John Mathias Jonathon C Scott, John Tehranian. The production of legal identities proper to states: The case of the permanent family surname. *COMPARATIVE STUDIES IN SOCIETY AND HISTORY*, 44, 2002.

[JJ20] Takayuki Mizuno Joomi Jun. Detecting ethnic spatial distribution of business people using machine learning. *Information*, 11:197, 2020.

[JJ21a] Takayuki Mizuno Joomi Jun. Analysis of ethnic homophily in international trade using a large-scale surname data. *The Review of Socionetwork Strategies*, 2021.

[JJ21b] Takayuki Mizuno Joomi Jun. Detecting ethnic linkages in economic networks using machine learning. *Big Data Analysis on Global Community Formation and Isolation*, pages 325–351, 2021.

[JL17] Miyoung Ko Donghee Choi Jaehoon Choi Jaewoo Kang Jinhyuk Lee, Hyunjae Kim. Name nationality classification with recurrent neural networks. *IJCAI*, pages 2081–2087, 2017.

[JQ11] Philippe Bourgois James Quesada, Laurie Kain Hart. Structural vulnerability and health: Latino migrant laborers in the united states. *Medical Anthropology*, 30:339–362, 2011.

[Ker11] Myleah Y. Kerns. North american women's surname choice based on ethnicity and self-identification as feminists. *NAMES-A JOURNAL OF ONOMASTICS*, 59:104–117, 2011.

[KLS04] L. K. Weiss H. Fakhouri W. Sakr-G. Kau R. K. Severson K. L. Schwartz, A. Kulwicki. Cancer among arab americans in the metropolitan detroit area. 2004.

[lana] wikipedia. https://www.wikipedia.org/. Accessed: 2023-01-02.

[lanb] World population review. https://worldpopulationreview.com/. Accessed: 2023-01-02.

[lis] listref. https://www.medey.com/wp-content/uploads/2020/10/China-Tariffs-Section-301.pdf. Accessed: 2023-01-02.

[LKK$^+$17] Jinhyuk Lee, Hyunjae Kim, Miyoung Ko, Donghee Choi, and Jaewoo Kang Jaehoon Choi. Name nationality classification with recurrent neural networks. pages 2081–2087. International Joint Conferences on Artificial Intelligence, 2017.

[Mat07] Pablo Mateos. A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place*, 13:243–263, 2007.

[Mat14] Pablo Mateos. *Names, ethnicity and Populations.* 2014.

[MG02] N. Singh S. S. Anand F. Raja-F. Mawji S. Yusuf M. Gupta, A. V. Doobay. Risk factors, hospital management and outcomes after acute myocardial infarction in south asian canadians and matched control subjects. *Cmaj*, 166:717–722, 2002.

[MHDLD13] A. Van Geystelen G. Defraene N. Vanderheyden K. Matthys T. Wenseleers M. H. D. Larmuseau, J. Vanoverbeke and R. Decorte. Low historical rates of cuckoldry in a western european human population traced by y-chromosome and genealogical data. *PROCEEDINGS OF THE ROYAL SOCIETY B-BIOLOGICAL SCIENCES*, 280, 2013.

[MNE08] P. A. Morrison P. Pantoja N. Lurie M. N. Elliott, A. Fremont. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research*, 43:1722–1736, 2008.

[MP11] Ana-Maria Popescu Marco Pennacchiotti. A machine learning approach to twitter user classification. *ICWSM*, 11:281–288, 2011.

[MT09] Arai Mahmood and Peter Skogman Thoursie. Labor market discrimination in delhi: Evidence from a field experiment. *JOURNAL OF COMPARATIVE ECONOMICS*, 37:14–27, 2009.

[ORB] Orbis. https://www.bvdinfo.com/en-gb/our-products/data/international/orbis. Accessed: 2023-01-02.

[RAB09] Martin Rosvall, Daniel Axelsson, and Carl T. Bergstrom. The map equation. *Eut.Phys.J.Special Topics*, 178(13), 2009.

[Rou87] Peter J Rousseeuw. Silhouettes.a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 1987.

[Sad20] Sara Sadhwani. Asian american mobilization: The effect of candidates and districts on asian american voting behavior. *Political Behavior*, pages 1–27, 2020.

[San00] Monica Sans. Admixture studies in latin america: From the 20th to the 21st century. 72, 2000.

[SFD13] K. J. Coleman C. Koebnick S. J. Jacobsen S. F. Derose, R. Contreras. Race and ethnicity data quality and imputation using us census data in an integrated health system: the kaiser permanente southern california experience. *Medical Care Research and Review*, 70:330–345, 2013.

[SJW15] Virginia M. Miller Stacey J. Winham, Mariza de Andrade. Genetics of cardiovascular disease: Importance of sex and ethnicity. *Atherosclerosis*, 241:219–228, 2015.

[SLS99] S. L. Glaser P. L. Horn-Ross D. W. West S. L. Stewart, K. C. Swallen. Comparison of methods for classifying hispanic ethnicity in a population-based cancer registry. *American Journal of Epidemiology*, 149:1063–1071, 1999.

[Tri02] Rauch J Trindade. Ethnic chinese networks in international trade. *The Review of Economics and Statistics*, 84, 2002.

[USC] Census. `https://www.census.gov/topics/population/genealogy/data.html`. Accessed: 2023-01-02.

[VIT16] Sneha Agarwal Vetle I. Torvik. *Ethnea–an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database*. 2016.

[WL13] Derek Ruths Wendy Liu. *What's in a name? using first names as features for gender inference in twitter*. 2013.

[YS19] Y. Wang J. Chen Y. Yuan H. E. Stanley Y. Shi, L. Li. Regional surname affinity: A spatial network approach. *American Journal of Physical Anthropology*, 168:428–437, 2019.