

A COMPUTATIONAL METHODOLOGY TO
MEASURE THE CULTURAL IDENTITY OF
COUNTRIES

CAROLINA COIMBRA VIEIRA

**A COMPUTATIONAL METHODOLOGY TO
MEASURE THE CULTURAL IDENTITY OF
COUNTRIES**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: PEDRO OLMO STANCIOLI VAZ DE MELO
COORIENTADOR: FABRÍCIO BENEVENUTO DE SOUZA

Belo Horizonte

Março de 2020

CAROLINA COIMBRA VIEIRA

A COMPUTATIONAL METHODOLOGY TO
MEASURE THE CULTURAL IDENTITY OF
COUNTRIES

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais, Instituto de Ciência Exatas, Departamento de Ciência da Computação. in partial fulfillment of the requirements for the degree of Master in Computer Science.

ADVISOR: PEDRO OLMO STANCIOLI VAZ DE MELO
CO-ADVISOR: FABRÍCIO BENEVENUTO DE SOUZA

Belo Horizonte

March 2020

© 2020, Carolina Coimbra Vieira.
Todos os direitos reservados.

Vieira, Carolina Coimbra.

V657c A Computational Methodology to Measure the
Cultural Identity of Countries [manuscrito] / Carolina
Coimbra Vieira. — 2020.
xviii, 64 f.; il.; 29cm.

Orientador: Pedro Olmo Stancioli Vaz de Melo.
Coorientador: Fabrício Benevenuto de Souza.

Dissertação (mestrado) — Universidade Federal de
Minas Gerais, Instituto de Ciência Exatas,
Departamento de Ciência da Computação.
Referências: f. 59-64

1. Computação — Teses. 2. Redes sociais on-line –
Teses. 3. Facebook (Recursos eletrônicos) – Teses.
4. Publicidade em Mídias — Teses. 5. Identidade
Cultural — Brasil — Teses. I. Melo, Pedro Olmo
Stancioli Vaz de. II. Souza, Fabrício Benevenuto de.
III. Universidade Federal de Minas Gerais, Instituto de
Ciências Exatas, Departamento de Ciência da
Computação. IV. Título.

CDU 519.6*22(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg
Lucas Cruz CRB 6ª Região nº 819.



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

A Computational Methodology to Measure the Cultural Identity of
Countries


CAROLINA COIMBRA VIEIRA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:


PROF. PEDRO OLMO STANCIOLI VAZ DE MELO - Orientador
Departamento de Ciência da Computação - UFMG


PROF. FABRÍCIO BENEVENUTO DE SOUZA - Coorientador
Departamento de Ciência da Computação - UFMG


PROF. RENATO MARTINS ASSUNÇÃO
Departamento de Ciência da Computação - UFMG


PROF. BERNARDO LANZA QUEIROZ
Departamento de Demografia - UFMG

Belo Horizonte, 27 de Março de 2020.

Resumo

Ao longo das diferentes ondas de migrantes a adoção das normas da sociedade anfitriã, bem como a propagação de suas próprias culturas trouxeram diversidade cultural ao país de destino. Embora seja fundamental para compreender as sociedades, medir a cultura e sua evolução tem sido uma meta complexa e elusiva, sobretudo devido à escassez e ao custo de obtenção de dados. Dessa forma, fontes de dados alternativas vem sendo cada vez mais utilizadas como fonte de dado substituindo ou complementando fontes de dado tradicionais. O uso de dados de redes sociais são um exemplo. O crescimento das redes sociais online nos últimos anos é impressionante. Somente o Facebook, a rede social mais popular, possui quase 2,5 bilhões de usuários ativos mensais. O número cada vez maior de usuários do Facebook representa novos clientes potenciais de empresas que pagam por espaços publicitários na rede social. De fato, a maior parte do faturamento das redes sociais online está concentrada em suas plataformas de marketing. Ao usar plataformas de anúncios de redes sociais online, um anunciante pode explorar publicidade de segmentação múltipla, o que permite selecionar usuários com características muito particulares, incluindo milhares de atributos demográficos, interesses e comportamentos. As redes sociais fornecem o ambiente ideal para inferir dados das populações online, uma vez que os usuários compartilham um grande número de informações pessoais, bem como sinais comportamentais, como curtidas e compartilhamentos de conteúdo de que gostam. Baseado nisso, aproveitamos as informações agregadas sobre os usuários fornecidas pela plataforma de publicidade do Facebook aos anunciantes para desenvolver uma metodologia para inferir a identidade cultural de países. Neste trabalho, propomos e desenvolvemos uma metodologia para não só identificar e caracterizar a identidade cultural como também medir a distância cultural entre países com base no interesse dos usuários por atributos culturais. Muitos aspectos culturais caracterizam as regiões em termos de atributos culturais, como roupas, música, arte e comida. Por isso, além da metodologia proposta, como exemplo de aplicação, apresentamos um estudo de caso focado em elementos culturais relacionados à culinária brasileira usando dados da Plataforma de Publicidade do Facebook. Através

deste estudo de caso, foi possível identificar os pratos típicos brasileiros que melhor se relacionam à identidade cultural do Brasil. Além disso, medimos a disseminação global da cultura alimentar brasileira entre países, explorando as preferências do usuário do Facebook por pratos típicos brasileiros. Os resultados mostram uma alta correlação entre a proporção de imigrantes brasileiros em cada país e a distância entre esses países e o Brasil em termos da distância cultural proposta. Por esse motivo, essa medida de distância pode complementar outras métricas de distância aplicadas a modelos do tipo gravidade, por exemplo, a fim de explicar o fluxo de pessoas entre os países.

Palavras-chave: Redes Sociais Online, Plataformas de Publicidade em Mídias Sociais, Identidade Cultural, Distância Cultural.

Abstract

Throughout the different waves of migrants, the adoption of the rules of the host society, as well as the propagation of their own cultures brought cultural diversity to the destination country. Although it is essential to understand societies, measuring culture and its evolution has been a complex and elusive goal, mainly due to the scarcity and the cost of obtaining data. Thus, alternative data sources are increasingly used as a data source, replacing or complementing traditional data sources. The use of social media data is an example. The growth of online social networks in recent years is impressive. Only Facebook, the most popular social network, has nearly 2.5 billion monthly active users. The growing number of Facebook users represents new potential customers from companies that pay for advertising space on the social network. In fact, most of the revenue from online social networks are concentrated on their marketing platforms. By using online social media ad platforms, an advertiser can explore micro-targeting advertising, which allows them to select users with very particular characteristics, including thousands of demographic attributes, interests, and behaviors. Social networks provide the ideal environment for inferring data from online populations, since users share a large number of personal information, as well as behavioral signals, such as likes and content shares they like. Based on this, we took advantage of the aggregated information about users provided by Facebook's advertising platform to advertisers to develop a methodology to infer the cultural identity of countries. In this work, we propose and develop a methodology to not only identify and characterize the cultural identity but also to measure cultural distance between countries based on users' interest in cultural attributes. Many cultural aspects characterize the regions in terms of cultural attributes, such as clothing, music, art, and food. Therefore, in addition to the proposed methodology, as an application example, we present a case study focused on cultural elements related to Brazilian cuisine using data from the Facebook Advertising Platform. Through this case study, it was possible to identify the typical Brazilian dishes that best relate to the cultural identity of Brazil. In addition, we measure the global spread of Brazilian food culture among

countries, exploring the Facebook user's preferences for typical Brazilian dishes. The results show a high correlation between the proportion of Brazilian immigrants in each country and the distance between those countries and Brazil in terms of the proposed cultural distance. For this reason, this distance measure can complement other distance metrics applied to gravity-type models, for example, in order to explain the flow of people between countries.

Keywords: Online Social Networks, Social Media Advertising Platforms, Cultural Identity, Cultural Distance.

List of Figures

3.1	Examples of interests probabilities distributions.	24
3.2	Matrix M''	25
3.3	Relationship between the entropy measure of the posterior distribution $H_{\pi_i^*}$ and the probability of the interest in the candidate country p^* . The x-axis represents the values for the probability p^* and the blue curve represents the correspondent value for $H_{\pi_i^*}$, represented by the y-axis, following the equation $H_{\pi_i^*} = \frac{-p^* \log(p^*) - (1-p^*) \log(1-p^*)}{p^*}$	30
3.4	Example of identification when an interest is part of the cultural identity of a country by varying the probability of one country, p_1 , and ϵ . We are considering a distribution over two countries where $p_2 = 1 - (p_1 + \epsilon)$. The green and the red markers represent when the interest is or is not part of the cultural identity of one of those countries, respectively.	31
4.1	Facebook Ads Platform.	37
5.1	Real population and Facebook audience in each country (log scale).	42
5.2	Proportion of interest in each country. All the interests are normalized by the audience in each country.	44
5.3	<i>Immigrant, Expat (Facebook) rankings, Geographic and Geographic weighted distance ranking.</i>	45
5.4	Comparison between <i>Immigrant</i> and Cosine distance rankings.	48
5.5	Comparison between the measure of interest entropy and entropy difference.	50
5.6	Relation between the entropy measure of the posterior distribution $H_{\pi_i^*}$ and the probability of the interest in Brazil p^*	51
5.7	Interest focus.	52

List of Tables

3.1	Example of the absolute audience data.	13
3.2	Example of the audience normalized by population.	13
3.3	Example of the z-score applied to the audience normalized by population.	15
3.4	Interest identified as part of the cultural identity of a country considering each example of interest probability distributions for each measure of cultural identity. The green markers represent interests that are part of the cultural identity of a country while the red markers represent the interests that could not be associated with the cultural identity of only one country.	26
4.1	Facebook audience for some typical Brazilian dishes.	38
5.1	Comparison with the <i>Immigrant ranking</i>	46
5.2	Comparison with the <i>Geographic weighted distance ranking</i>	46
5.3	Comparison with the <i>Immigrant ranking</i> . Considering only the 7 interests typical from Brazil according to entropy difference.	53
5.4	Comparison with the <i>Geographic weighted distance ranking</i> . Considering only the 7 interests typical from Brazil according to entropy difference.	53

Contents

Resumo	ix
Abstract	xi
List of Figures	xiii
List of Tables	xv
1 Introduction	1
2 Related Work	5
2.1 Cultural distance	5
2.1.1 Food as a proxy indicator of culture	6
2.1.2 Online social media data	6
2.2 Cultural identity	7
2.3 Facebook data	8
3 Methodology	11
3.1 Problem definition	11
3.2 Input	12
3.3 Cultural distance	13
3.3.1 Problem definition	14
3.3.2 Input: z-score normalization	14
3.3.3 Measures of distance	16
3.3.4 Metrics to compare rankings	18
3.4 Cultural identity	20
3.4.1 Problem Definition	21
3.4.2 Input: interest probability distribution	22
3.4.3 Measures of identity	24

4	Gathering Facebook Data	35
4.1	Background on Facebook Ads	35
4.2	Collecting Facebook Ads data	36
4.3	Privacy and Limitations	38
5	The Case of Brazilian Cuisine	41
5.1	Data description	41
5.2	Brazilian cultural distance	42
5.2.1	Baseline data	43
5.2.2	Comparison between rankings	46
5.3	Brazilian cultural identity	49
5.4	Discussion	53
6	Conclusion and Future Work	55
6.1	Conclusion	55
6.2	Future Work	56
	Bibliography	59

Chapter 1

Introduction

Despite the differences between cultures, many populations even living in different places share some cultural preferences. While key for understanding societies, characterizing the culture and measuring the cultural distance between countries has been a complex and elusive goal. Different waves of migrants have adopted norms of the host society while also bringing cultural diversity to the destination country. According to the World Migration Report 2018 [of Migration, 2017] the vast majority of people migrate voluntarily, due to studies, economic, and family reasons. However, many other factors can contribute to the increasing number of people migrating voluntary or involuntarily, such as diseases¹, conflicts², poverty³, and disasters⁴.

In order to study international trade and migration, gravity models are one type of model ordinarily applied by social scientists [Isard, 1954, Cohen et al., 2008, Massey et al., 1993]. The most important element in those models is the distance measure that characterizes the attraction between regions. In other words, in those models, the attraction is proportional to the similarity and inversely proportional to the distance between countries. The distance or similarity between two countries can be measured by administrative and political distance, geographical distance, economic distance, and also in terms of cultural distance [Ghemawat, 2001]. All of these distance metrics complement other metrics of distance being incorporated in models to explain the flows of people between countries.

¹<https://www.reuters.com/article/us-china-health-repatriation/china-says-will-repatriate-overseas-citizens-if-needed-due-to-coronavirus-idUSKBN2001FT>

²<https://www.pbs.org/wgbh/frontline/article/numbers-syrian-refugees-around-world/>

³<https://www.mercycorps.org/blog/quick-facts-venezuela-crisis>

⁴<https://weather.com/news/news/2018-12-20-puerto-rico-migration-new-york-population-maria-census>

Focusing on cultural distance, many aspects may help us to culturally characterize regions before calculating distance metrics among them, such as preferences for clothes, music, art, and food [Recchi and Favell, 2019]. The cuisine of a country, for example, reflects its history, while the influx of immigrants from many foreign nations develops a rich diversity in food preparation throughout the country⁵. As such, cuisine can be used as a proxy indicator of culture in a country and the number of people interested in typical food [Sibal, 2018, Boutaud et al., 2016, Almerico, 2014], can be used to estimate the strength of that culture inside the region.

When we think about tacos, for example, it is clear that we are talking about Mexican cuisine. The same happens when we talk about sushi, a typical Japanese dish. But, for some food like bread and rice, it is not easy to find only one specific region characterized by this kind of food. Some dishes clearly correspond to the cultural identity of a specific country. Others can be associated with more than one country. Notice that if we change to another context, for example, music genre, the same pattern happens. Think about samba: which country comes to your mind? And when you think about rock? In the first case, it is clear the region we are referring to. When this match between an interest representing a cultural attribute and a region happens we say that there is a cultural relationship between them. If we focus on selecting those cultural aspects to be used as a proxy of culture in a country, there is no automatic way to classify them as being from a specific country. In other words, it is not trivial to characterize a country in terms of cultural aspects. Mainly because of the amount of data available nowadays, it is important to develop new methodologies to classify automatically those cultural aspects giving the right importance to the owner country.

As one of the most expressive data sources, we can consider the Online Social Networks such as Facebook. The key role in the business model of these online social networks is advertising which underpins much of the Internet's economy. One of the keys for the success of online social networks advertising platforms is the vast possibilities to reach users by providing a list of personally identifiable information (name, phone number, email, etc) or by configuring targeting options from a huge list of fine-grained attributes such as race, income level, interests, and behaviors. By using Facebook data, we have access to the data provided by the users from the biggest social network. Facebook users register their friends, events, communities, and interests when they interact with videos, pictures, and pages. We hypothesized that people from similar cultures will have similar interests and they will be expressed through the huge amount of data available on the Facebook Advertising Platform, Facebook Ads.

⁵<https://freelymagazine.com/2017/01/07/what-food-tells-us-about-culture/>

Using social media data to study the cultural identity and measure the cultural distance between countries is not completely new. The use of the data provided by Facebook Ads, for example, has been recently increasing in many different fields. Stewart et al. [Stewart et al., 2019], discuss the cultural diffusion or the cultural assimilation of immigrants by considering the taste by music gender. Our work makes use of the same data source but focuses on the cuisine as a marker of culture. Other authors explore different social media data for similar purposes. Considering the cuisine around the world, Silva et al. [Silva et al., 2014] identify cultural boundaries by analyzing food and drink habits in Foursquare. They identify cultural boundaries and similarities across populations by clustering them based on the analysis of Foursquare check-ins. Similarly to us, they calculate the cultural distance between two countries by considering the habits of Foursquare users. Although there are some differences in the way we deal with the problem. First, the data source is different. Second, the data provided by Foursquare refers to check-ins. Because of this, the interest in some kind of food is restricted to the users' check-in in those places.

In this scenario, we propose a method that explores the Facebook Advertising Platform to infer the number of people interested in some cultural aspects. We are interested in (1) quantifying the cultural distance between countries and (2) identifying the cultural identity of countries by exploring Facebook users' preferences through the data retrieved from the Facebook Advertising Platform.

To characterize and measure the cultural distance between countries, we develop a new methodology for exploring the data provided by online social networks on advertising platforms. By using z-score normalization, we create a vectorized representation of countries in order to compare them to each other. While evaluating the cultural distance between countries, we explore several measures of distance and compare them in the context of cultural affinities. To decide which interest is part of the cultural identity of each country, we made use of an approach that relies on information theory and is based on a measure of entropy.

Particularly, we validate our methodology by employing it in a case study where typical Brazilian dishes are used as a proxy of how the Brazilian culture is consumed across various countries in the world. From this, we also measure the cultural distance between Brazil and the countries most preferred by Brazilian immigrants. To do that, we selected 20 typical Brazilian dishes according to some websites like Wikipedia and collected the data on interests for these dishes from the Facebook Advertising Platform (Facebook Ads). Using the number of Facebook users interested in certain typical Brazilian food in each country enables us to represent such interests in each country via a vector. This allows us to evaluate, in a natural way, the distance between pairs

of countries. We also apply the methodology to identify dishes that in fact represent the identity of the country.

The main contributions of this work are:

1. **A new measure of cultural distance**
2. **Methods to identify the cultural identity of countries**
3. **An analysis of the Brazilian cuisine as a proxy of cultural identity**

By using the methodology to measure cultural distance between countries, it is possible to apply a simple clustering technique, using this cultural distance measure, to draw cultural boundaries across countries. As a result, a cultural map of the world can be generated. This measure of distance is a new way to express the attraction between countries in terms of cultural distance and it can be incorporated into gravity-type models to explain the flows of people between countries. The methodology to identify the relationship between interests and countries is important to characterize countries and helps to identify the main aspects of a culture. With the methodology we propose, it is possible to infer the cultural attributes of a country or region given only the user's interests. Finally, these approaches might be useful not only for economic purposes but also to support existing and novel marketing and social applications. This thesis is organized as follows:

Chapter 2 presents the background and related works.

Chapter 3 details the methodology used to represent the data and describes the data normalization process, including the methodology to measure the cultural distance between countries and to characterize countries in terms of cultural interests.

Chapter 4 details the Facebook data and how to collect the data provided by the Facebook Advertising Platform.

Chapter 5 presents the case study of how our methodology can be used to characterize Brazil in terms of interests regarding Brazilian cuisine. We also measure the cultural distance from Brazil to other countries by using the methodology proposed in Chapter 3.

Finally, in Chapter 6 discusses the results and offer additional comments about the applicability of the results on Gravity-type models for migration. We also discuss our findings, draw final conclusions, and future work.

Chapter 2

Related Work

In this chapter, we review related work. In Section 2.1 we discuss several papers related to the measure of the distance between countries, specifically, focusing on the cultural distance. This section includes related work about the use of food as a proxy indicator of the culture and the use of online social media data in the measure of cultural distance. Afterward, in Section 2.2 we review the current state-of-the-art in cultural identity. Finally, in Section 2.3 we review the current works that make use of Facebook data.

2.1 Cultural distance

The study of international migration and the development of models to explain and to predict flows of people between countries is not new [Massey et al., 1993]. One of the most important methods is based on the gravity-type models. Cohen et al. [Cohen et al., 2008] developed an algorithm to project future numbers of international migrants from any country or region to any other. Basically, the variables considered by the model include the population and area of origins and destinations of migrants and the geographic distance between origin and destination. Other researchers point out other distance measures beyond the geographic distance, like administrative and political distance, economic distance, and also in terms of cultural distance [Ghemawat, 2001].

Several papers attempted to classify cultural aspects to compare countries in terms of their cultural distance. In [Ghemawat, 2001], the authors list a few types of distances that can be considered when comparing regions and the impacts that each of these distances has, mainly on the financial sphere: they found cultural distance to be one of the most important factors. Another line of literature outlines many

characteristics that may represent a culture of a country and used these factors to cluster countries according to their similarity. Santis et al. [De Santis et al., 2015] presented a new method for evaluating the relative distance between any two countries by using individual data provided by the World Value Survey (WVS).

In the next section, we present other attributes of daily life that can be used as a proxy of culture in a country. We focus on presenting works that discuss food as a cultural characteristic shared by individuals since we use food as a proxy of the Brazilian cultural identity in the case study, presented in Chapter 5.

2.1.1 Food as a proxy indicator of culture

In anthropology and sociology, the study of culture can be examined considering a multitude of aspects of our daily life, such as the clothes we wear, the music we listen to, and the food we eat [Recchi and Favell, 2019]. Food studies are an established interdisciplinary field that recognizes the centrality of food for cultural practices and cultural identity. Considering the dishes from a country or region, for example, it is possible to approximate cultural distance by characterizing the preferences for local foods [Sibal, 2018, Boutaud et al., 2016, Almerico, 2014].

Several works, such as [Sibal, 2018], explore the idea that food communicates our culture and mechanisms by which we relate food to our cultural identities, while others revealed that people and societies can be discriminated against by their food and cultural habits [Boutaud et al., 2016, Almerico, 2014]. Sibal [Boutaud et al., 2016] focuses on showing the diversity and similarities between people, cultures, and food. Similarly, Almerico [Almerico, 2014] presented an interdisciplinary study that observes the intricate relationships between food, culture, and society from the sociological perspective.

Notice that most parts of the works regarding the relationship between food and culture presented in this section are more theoretical and does not explore online data. In the next section, we present some papers that explore online social media data to study aspects related to food, culture, and cultural distance between countries.

2.1.2 Online social media data

According to [Sajadmanesh et al., 2017], food and nutrition occupy an increasingly prevalent space on the web. Dishes and recipes shared online provide an invaluable mirror into culinary cultures and attitudes around the world. By exploring worldwide culinary habits on the web, the authors confirm the strong effects of geographical and

cultural similarities on recipes, health indicators, and culinary preferences between countries. Other authors focus on spatial and temporal patterns of online food preferences [Wagner et al., 2014, West et al., 2013]. Abbar et al. [Abbar et al., 2015] build a model to predict county-wide obesity and diabetes statistics based on a combination of demographic variables and food names mentioned on Twitter.

Since the link between typical food and culture of a country has been established, several papers attempted to classify those aspects in order to compare countries in terms of their cultural distance. Silva et al. [Silva et al., 2014] identify cultural boundaries by analyzing food and drink habits in Foursquare. They identify cultural boundaries and similarities across populations by clustering them based on the analysis of Foursquare check-ins. Similarly to us, they calculate the cultural distance between countries by considering the habits of Foursquare users. Although there are some differences in the way we deal with the problem. First, the data source is different. Second, the data provided by Foursquare refers to check-ins, because of this, even the people have an interest in some kind of food, but do not have a check-in in those places, this information will not be available.

To the best of our knowledge, this is the first work that uses Facebook data to explore the cultural identity of countries to measure the cultural distance between them. In addition to the methodology to measure cultural distance between countries, part of the methodology presented in Chapter 3, refers to characterize countries in terms of cultural attributes. In the next section, we present papers that discuss the state-of-the-art for identifying cultural identity.

2.2 Cultural identity

Because of globalization, it is difficult to characterize a country in terms of cultural attributes. Usually, when cultural psychology work tries to compare countries on various dimensions, essentially they use the same approach: summarizing a large number of variables into principal components or factors, then standardizing these scores, and comparing them across countries [Beugelsdijk and Welzel, 2018]. The Hofstede model¹ of national culture consists of six dimensions. The cultural dimensions represent independent preferences for one state of affairs over another that distinguish countries (rather than individuals) from each other. To the best of our knowledge, this is the first work that uses Facebook data to characterize the cultural identity of countries.

¹<https://www.hofstede-insights.com/models/national-culture/>

Other approaches adopted in studies, mainly in the field of psychology, are questionnaires and interviews. Bhugra et al. [Bhugra et al., 1999] describe the key concepts of cultural identity such as religion, attitudes to the family, leisure activities, rites of passages, food, and language. They discuss the difficulty in measuring the cultural identity and the various measurements used during the interviews illustrate the need of having a multifaceted instrument in measuring cultural identity among Asians. Cantarero et al. [Cantarero et al., 2013] also identify food as greatly influenced by cultural identity by performing a qualitative and quantitative analysis in the Comunidad Autónoma de Aragón, Spain. The research methods include focus groups, in-depth interviews, participant observation, and a questionnaire. Regarding the research outcome, they found that people prefer to consume foods that are symbolically associated with their own culture, in order to reinforce their sense of belonging. Although this study has been carried out in Aragón, the results can be generalized to other areas.

Until now, we discuss previous works related to our methodology to measure cultural distance and to identify the cultural identity of countries, presented in Chapter 3. However, part of this thesis consists of the use an online social media data, specifically the data provided by the Facebook Advertising Platform. In the next section, we present previous papers that make use of this data in different applications, such as migration.

2.3 Facebook data

The key role in the business model of these online social networks is advertising [Weber et al., 2018]. Guha et al. [Guha et al., 2010] conducted one of the first studies that analyze methodologies for ad networks. They show how location, user demographics, and interests, and sexual-preference affect Facebook Ads. The Facebook Advertisement Platform allows advertisers to target users by demographic characteristics including age, location, and interests as determined by their profile information and provides the number of users who fall under the selected categories. Because of this, the data provided by the Facebook Advertising Platform is becoming popular especially in demographic studies [Alburez-Gutierrez et al., 2019].

Chunara et al. [Chunara et al., 2013] developed one of the first studies by using data provided by the Facebook Advertisement Platform. They examined activity and sedentary related interest categories from Facebook that have, in other traditional studies of the social environment, been positively or negatively related to obesity.

Since before, the use of the marketing tool provided by Facebook, Face-

book Ads, to access and collect data is increasing in many different fields. It was employed in different contexts, such as in tracking health conditions [Araujo et al., 2017, Mejova et al., 2018b, Rampazzo et al., 2018], predicting crimes [Fatehkia et al., 2019], gender inequalities [Garcia et al., 2018, Fatehkia et al., 2018, Mejova et al., 2018a, Gil-Clavel and Zaghene, 2019], political science [Ribeiro et al., 2018, Ribeiro et al., 2019, Silva et al., 2020], and to study migration [Zaghene et al., 2017, Pötzschke and Braun, 2017, Dubois et al., 2018, Spyrtos et al., 2018, Spyrtos et al., 2019, Palotti et al., 2020], relationships between immigrant communities [Herdağdelen et al., 2016] and migrant assimilation [Dubois et al., 2018, Stewart et al., 2019].

Recent papers explore the Facebook data to characterize refugees and populations living in inequality situations [Rama et al., 2020, Palotti et al., 2020]. Rama et al. [Rama et al., 2020] examine the usefulness of the Facebook Advertising Platform, which offers a digital “census” of over two billion of its users, in measuring potential rural-urban inequalities in Italy. Palotti et al. [Palotti et al., 2020] use Facebook data as an additional data source for monitoring the ongoing crisis in Venezuela. They estimate and validate national and sub-national numbers of refugees and migrants and break-down their socio-economic profiles.

To the best of our knowledge, we developed the first study using this data to measure the cultural distance between countries [Vieira et al., 2020] that will be presented in Chapter 5. Stewart et al. [Stewart et al., 2019] is the most similar paper discussing cultural diffusion. However, they analyze the cultural assimilation of immigrants while our work compares countries at a high level by analyzing the population interests in each country as a whole. We do not analyze separately the interests of immigrants and natives. Nevertheless, we are aware that this analysis is important for future research in terms of quantifying the extent to which immigrants affect the culture of natives.

Finally, it is important to point out that even with the nature of the service constructed as a “black box”, and a slew of biases [Cesare et al., 2018, Araujo et al., 2017], including the algorithmic bias in extracting user attributes, all these works provide evidence that Facebook Ads data can indeed be used as a source of information for the study of computational social science.

In Chapter 4, we discuss the process of collecting Facebook data in order to have significant data to be used as an input of the methodologies presented in Chapter 3. The methodologies to identify the cultural identity of countries as well as to measure the cultural distance between countries are exemplified in Chapter 5, where we present a case study. The case study where typical Brazilian dishes are used as a proxy of the Brazilian culture identity is used to explain the Facebook data collection and to

validate our methodology.

Chapter 3

Methodology

In this thesis, we explore the process of measuring the cultural distance between countries and to infer the cultural identity of countries. The methodology adopted is described in the remainder of this chapter, which is organized as follows. First, we define the problem formally. Then, we describe the data and the normalization process. Next, we describe our methodology to represent countries via vectors and to measure the cultural distance between them. Finally, we detail a novel methodology to infer the cultural identity of countries, including examples and discussing some limitations of our approach.

3.1 Problem definition

The input of the problem is given by triples $t = (c, i, a)$ where c corresponds to a country, i corresponds to an interest and a corresponds to the audience. We define by **audience** the number of people who lives in c and is interested in i . Interests represent cultural aspects that may help us to culturally characterize regions, such as preferences for clothes, music, art, and food. So, given the data that represents the interests of people in each country, we can define our two problems as follows:

- **How to measure the cultural distance between countries?**
- **How to identify cultural aspects that make up the identity of countries?**

As a first step to model the problem, we detail the process of structuring the data in Section 3.2. Then, the methodology to measure the cultural distance between countries and to infer the cultural identity of countries are described in Sections 3.3 and 3.4, respectively.

3.2 Input

Given the input of the problem as triples $t = (c, i, a)$, we can structure them as a matrix M where each cell $M_{i,c}$ represents the number of people interested in i from the country c . In other words, each cell represents the audience $A(i, c)$ in c who are interested in i .

The audience in each country can vary a great deal across countries. Countries with small populations, for example, tend to have lower values for the majority of the interests. On the other hand, countries with large populations are more likely to have large audiences. This bias created by the discrepancy in the size of populations makes the analyses unfair since countries with large populations dominate the audience over an interest.

During the process to find interests that are part of the cultural identity of a country the main goal is: given an interest, link it to a country. Basically, for each interest, we want to identify the country that dominates the audience considering all the countries. Because of this, countries with large populations tend to present high absolute audiences, even when the proportion of the population interested in them is not high.

Thus, to remove the bias by considering the absolute number of people interested in i from the country c , we normalize the audience in each interest by the population in each country. We use these proportions to construct a new matrix M' to represent the normalized data. More formally, given the population $A(c)$ of a country c and the audience $A(i, c)$ in c who are interested in i , the normalization is given by:

$$M'_{i,c} = \frac{A(i, c)}{A(c)} \quad (3.1)$$

As shown by Equation 3.1, each cell $M'_{i,c}$ represents the proportion of the population in c who are interested in i .

Consider the example of three interests, $I1$, $I2$, and $I3$, and the number of people interested in them from five countries, $C1$, $C2$, $C3$, $C4$, and $C5$, represented by the matrix M , where the rows correspond to interests and the columns represent the countries. The absolute audience of each interest in each country is represented by a matrix M in Table 3.1. Each one of the five positions in the row corresponds to the audience in the corresponding country. In a first look, the interest $I1$ seems to be more popular in country $C1$, interest $I2$ seems to be popular in country $C4$, and interest $I3$ more popular in country $C5$.

Consider now that the population in $C1$, $C2$, $C3$, $C4$, and $C5$ is, respectively,

$$M = \begin{array}{c} \\ \\ \\ \end{array} \begin{array}{ccccc} & C1 & C2 & C3 & C4 & C5 \\ I1 & 1000 & 10 & 1 & 100 & 500 \\ I2 & 100 & 100 & 10 & 1000 & 10 \\ I3 & 10000 & 200 & 100 & 50000 & 1000 \end{array}$$

Table 3.1: Example of the absolute audience data.

$$M' = \begin{array}{c} \\ \\ \\ \end{array} \begin{array}{ccccc} & C1 & C2 & C3 & C4 & C5 \\ I1 & 0.01 & 0.02 & 0.001 & 0.0001 & 0.33 \\ I2 & 0.001 & 0.2 & 0.01 & 0.001 & 0.006 \\ I3 & 0.1 & 0.4 & 0.1 & 0.05 & 0.66 \end{array}$$

Table 3.2: Example of the audience normalized by population.

100000, 500, 1000, 1000000 and 1500. By transforming the absolute audience to the proportion of the audience considering the population in each country, each cell of the matrix M will be updated according to Equation 3.1. The new matrix, M' , with the data normalized is shown by Table 3.2. Now, the interest $I2$ seems to be more popular between the proportion of the audience who lives in the country $C2$. This change also happens with interest $I1$ that now seems to be much more popular in country $C2$, 20% of its population is interested in $I2$. Only the interest $I3$ maintains the popularity in the proportion of the population in country $C5$.

After the process of structuring and normalizing the data, the representation is given by a matrix where the rows correspond to interests and the columns represent countries. If we consider each column of the matrix as a representation of each country, we have a vector where each position corresponds to the proportion of its population who are interested in each interest. We can also make use of this representation to compare countries with each other as we will discuss in Section 3.3 and to identify the cultural identity of countries as presented in Section 3.4.

3.3 Cultural distance

In order to answer the question **How can we measure the cultural distance between countries?**, we dedicate this section to model this specific problem given the input described in the previous section. We argue that a proper methodology for measuring cultural distance must contain:

1. The definition of cultural distance in our context which will be presented in Section 3.3.1.

2. A description of the process of data normalization to create vector representations for each country which will be described in Section 3.3.2.
3. Experiments using several different measures of distance to compare the countries' vectors which will be presented in Section 3.3.3.
4. Complementary and robust metrics to compare rankings. In Section 3.3.4 the main important metrics to compare rankings are described.

3.3.1 Problem definition

Cultural distance: Cultural distance is a measure of how distant two regions are considering cultural aspects. Usually, the cultural distance dimensions refer to values on which societies or nations differ, such as power distance and uncertainty avoidance. Cultural distance measures refer to the operationalization of these dimensions, often into a single score, which can be used to estimate cultural distance¹.

Cultural distance dimensions refer to national or societal values on which nations or societies tend to differ. Usually, cultural distance dimensions include, among others, power distance, uncertainty avoidance, collectivism, and assertiveness [Beugelsdijk and Welzel, 2018], based on the answers given to questionnaires on cultural values. Based on cultural distance dimensions, cultural distance measures are constructed by aggregating them into a single equation. Such measures can take the form of compound indices that bundle distance scores for individual cultural dimensions. One of the most commonly applied cultural distance measures is KS-index [Tung and Verbeke, 2010].

In our context, we are interested in measuring the distance between countries by considering the whole vector of interests that represents each country. In order to create the vector representation, the next section discuss the input data we have and how we create the vectorized representation.

3.3.2 Input: z-score normalization

Section 3.2 described the input we will consider for this problem. At the end of this section, the input is represented by a matrix where the rows correspond to interests and the columns represent countries. Each column corresponds to a vector representation of a country and each cell in this vector corresponds to the proportion of the population who are interested in each interest.

¹https://research-api.cbs.dk/ws/portalfiles/portal/58442485/ronni_kj_rhede.pdf

$$M'_z =$$

	C1	C2	C3	C4	C5
I1	-0.482	-0.406	-0.551	-0.558	1.997
I2	-0.546	1.998	-0.431	-0.546	-0.474
I3	-0.699	0.577	-0.699	-0.901	1.70

Table 3.3: Example of the z-score applied to the audience normalized by population.

Although the normalized representations by the population of the country, the difference in the popularity of interests can make the comparison of two representations be biased toward more popular interests. If we consider each column as a vector representing the country interests, these unbalanced distributions bias the distance measurement between two countries by the most popular interests.

Considering the example in Table 3.2, country $C1$ is represented by the vector $[0.01, 0.001, 0.1]$ and country $C2$ represented by the vector $[0.02, 0.2, 0.4]$. The absolute difference between each position of the vectors is given by the vector $[0.01, 0.199, 0.3]$. The maximum difference appears in the third position, corresponding to the interest $I3$, which is the interest most popular. By this example, we see that the difference between other interests that have a small proportion of the audience will not have the same importance as the interests most popular in measures of distance.

To avoid the problem of unbalanced distributions due to the presence of a dominant interest, the z-score normalization can be used to smooth the distribution and give the same importance for all interests. Equation 3.2 shows the formula for calculating the z-scores:

$$zscore(M'_{i,c}) = \frac{M'_{i,c} - mean(M'_i)}{std(M'_i)} \quad (3.2)$$

where M'_i is the vector that contains $M'_{i,c}$ for each country c .

Basically, the mean is subtracted from the score for each interest, normalized by the proportions of the audience in each country, $M'_{i,c}$, and divided by the standard deviation of the proportions for that interest in all the countries. As a result, each value now represents the extent measured in standard deviations to which an interest in a certain country deviates from the mean of a typical distribution. So, after the z-score normalization, each country is represented by a vector that can be compared to each other.

After the z-score normalization applied to the example in Table 3.2, the countries can be represented by the columns of the matrix M'_z , as shown in Table 3.3. Notice that, in this example, each country is represented by a vector of interests.

Given the vectorized representation for each country, in the next section, we

discuss the measures of distance that can be used to compare the distance between countries.

3.3.3 Measures of distance

There are many different metrics to measure the distance between vectors. As one of the most popular, we have Euclidean distance, Cosine distance, and Earth Mover's distance. Also, some measures of error can be used as a measure of the distance between two vectors by comparing the difference between each position in both vectors.

In this section, we will present some of the most important measures of distance between vectors. For all definitions, we assume that we are comparing two countries represented via vectors, \mathbf{c}_1 and \mathbf{c}_2 .

Euclidean distance:

Euclidean distance² is a measure of the true straight line distance between two points in Euclidean space as shown by Equation 3.3. Essentially, it measures the length of a segment that connects two points.

$$Euc(\mathbf{c}_1, \mathbf{c}_2) = \|\mathbf{c}_1 - \mathbf{c}_2\|_2 = \sqrt{\sum_{i=1}^N (c_{1i} - c_{2i})^2} \quad (3.3)$$

Euclidean distance is the most common distance for machine learning algorithms. However, there are some situations where Euclidean distance will fail to give us the proper metric, especially when the number of dimensions increases. In those cases, we will need to make use of different distance functions.

Cosine distance:

Cosine distance³ is a measure of the angle between two vectors as shown by Equation 3.4.

$$Cos(\mathbf{c}_1, \mathbf{c}_2) = 1 - \frac{\mathbf{c}_1 \cdot \mathbf{c}_2}{\|\mathbf{c}_1\|_2 \|\mathbf{c}_2\|_2} = 1 - \frac{\sum_{i=1}^N (c_{1i} \cdot c_{2i})}{\sqrt{\sum_{i=1}^N c_{1i}^2} \sqrt{\sum_{i=1}^N c_{2i}^2}} \quad (3.4)$$

In order to calculate it, we need to measure the cosine of the angle between two vectors. Then, cosine distance returns the normalized dot product of them. A

²<https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.spatial.distance.euclidean.html>

³<https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.spatial.distance.cosine.html>

normalized vector is a vector in the same direction but with norm 1. The dot product is the operation in which two equal-length vectors are multiplied resulting in a single scalar. Cosine distance is very useful when we are interested in the orientation but not the magnitude of the vectors. Two vectors with the same orientation have a cosine distance of 0. Two vectors at 90° have a distance of 1. Two vectors diametrically opposed having a distance of -1. All independent of their magnitude. This metric is a measurement of orientation and not magnitude. The Euclidean distance between two vectors can be higher even when the angle between them is small because they are pointing in the same direction.

Earth Mover’s distance:

This distance is also known as the earth mover’s distance⁴, since it can be seen as the minimum amount of “work” required to transform \mathbf{c}_1 into \mathbf{c}_2 , where “work” is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved as shown by Equation 3.5.

$$EMD(\mathbf{c}_1, \mathbf{c}_2) = \inf_{\pi \in \Gamma(\mathbf{c}_1, \mathbf{c}_2)} \int_{R \times R} |x - y| d\pi(x, y) = \int_{-\infty}^{+\infty} |\mathbf{C}_1 - \mathbf{C}_2| \quad (3.5)$$

where $\Gamma(\mathbf{c}_1, \mathbf{c}_2)$ is the set of (probability) distributions on $R \times R$ whose marginals are \mathbf{c}_1 and \mathbf{c}_2 on the first and second factors respectively. \mathbf{C}_1 and \mathbf{C}_2 are the respective Cumulative Distribution Functions (CDFs) of \mathbf{c}_1 and \mathbf{c}_2 .

Mean absolute error:

A direct approach to compare \mathbf{c}_1 and \mathbf{c}_2 is through the average of the errors between each cell of \mathbf{c}_1 and the corresponding cell in \mathbf{c}_2 , defined by Equation 3.6.

$$MAE(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{N} \sum_{i=1}^N |c_{1i} - c_{2i}| \quad (3.6)$$

Relative error:

Another approach is the average of the relative errors between each cell \mathbf{c}_1 and the corresponding cell in \mathbf{c}_2 , defined by Equation 3.7.

$$RE(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{N} \sum_{i=1}^N \frac{|c_{1i} - c_{2i}|}{c_{1i}} \quad (3.7)$$

⁴https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html

More important than measure the cultural distance between countries is to compare the results given by our measure of cultural distance with other existing measures. Basically, for each country, we can measure the cultural distance to others. The list of countries compared can be sorted by the cultural distance and expressed by a ranking. In the next section, we present some measures to compare rankings that can be used in order to compare and correlate our results with others in the literature.

3.3.4 Metrics to compare rankings

After calculating the distance between countries using one of the measures presented in Section 3.3.3, we can generate rankings sorted by the most similar countries in terms of culture to a target country. The rankings can also be compared with other rankings generated by other sources, such as official reports and works in literature. Ideally, we would like to compare the cultural ranking with different types of rankings that express not only a relation between cultural aspects but also geographic distance and social similarities given by the number of immigrants, for example.

There are several measures proposed to compare rankings. In this section, we will present four of them. For all definitions, we assume that we are comparing two rankings with country names, \mathbf{R}_1 and \mathbf{R}_2 . We apply these measures during our rankings comparison in the Case Study we present in Chapter 5.

Kendall tau⁵:

This metric uses Kendall's tau rank correlation coefficient [Knight, 1966], defined by the Equation 3.8.

$$K\tau(\mathbf{R}_1, \mathbf{R}_2) = \frac{C(\mathbf{R}_1, \mathbf{R}_2) - D(\mathbf{R}_1, \mathbf{R}_2)}{\sqrt{C(\mathbf{R}_1, \mathbf{R}_2) + D(\mathbf{R}_1, \mathbf{R}_2) + T(\mathbf{R}_1)}\sqrt{C(\mathbf{R}_1, \mathbf{R}_2) + D(\mathbf{R}_1, \mathbf{R}_2) + T(\mathbf{R}_2)}}, \quad (3.8)$$

where $C(\mathbf{R}_1, \mathbf{R}_2)$ is the number of concordant pairs, $D(\mathbf{R}_1, \mathbf{R}_2)$ is the number of discordant pairs, $T(\mathbf{R}_1)$ and $T(\mathbf{R}_2)$ are the number of ties only in \mathbf{R}_1 and \mathbf{R}_2 , respectively. Values close to 1 indicate strong agreement, values close to -1 indicate strong disagreement.

Spearman R⁶:

⁵<https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.kendalltau.html>

⁶<https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.spearmanr.html>

This metric calculates a Spearman rank-order correlation coefficient [Zwillinger and Kokoska, 1999], defined by the Equation 3.9.

$$Spearmanr(\mathbf{R}_1, \mathbf{R}_2) = \frac{cov(\mathbf{R}_1, \mathbf{R}_2)}{\sigma_{\mathbf{R}_1} \sigma_{\mathbf{R}_2}}, \quad (3.9)$$

where $cov(\mathbf{R}_1, \mathbf{R}_2)$ is the covariance of the rank variables and $\sigma_{\mathbf{R}_1}$ and $\sigma_{\mathbf{R}_2}$ are the standard deviations of the rank variables.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic relationship.

Weighted tau⁷:

This metric computes a weighted version of Kendall's tau rank correlation coefficient [Vigna, 2015].

This measure is important when we are interested in giving more importance to the first elements in the ranking. Hence, we decided to consider the measure of correlation that allots more weight to the top elements in the rank. The weight is mapped from non-negative integers (zero representing the most important element, the first in the ranking) to a non-negative weight, given by a hyperbolic weighing. The hyperbolic weighting maps the position of each element in rank r to a weight $\frac{1}{r+1}$. Because of this, the first element ($r = 0$) has a weight equal to 1, the second, $\frac{1}{2}$, and so on.

Jaccard similarity⁸:

This metric measures the coefficient of similarity between finite sample sets, given by Equation 3.10.

$$Jaccard(\mathbf{R}_1^*, \mathbf{R}_2^*) = \frac{|\mathbf{R}_1^* \cap \mathbf{R}_2^*|}{|\mathbf{R}_1^* \cup \mathbf{R}_2^*|}, \quad (3.10)$$

where $|\mathbf{R}_1^* \cap \mathbf{R}_2^*|$ and $|\mathbf{R}_1^* \cup \mathbf{R}_2^*|$ represent, respectively, the size of the intersection and the size of the union of the sample sets.

Unlike the other metrics, the Jaccard similarity does not consider the order with the elements appear in the ranking. The Jaccard coefficient is defined as the size of

⁷<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.weightedtau.html>

⁸https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html

the intersection divided by the size of the union of the sample sets. Note that, in this measure, the structure does not matter, only the fact if the element appears in the ranking or not. In our case, all the rankings have the same elements, because of this, if we consider the rankings as a whole, the Jaccard coefficient will be equal to 1. So, when we use this measure, we select only the first k elements, \mathbf{R}_1^* and \mathbf{R}_2^* , of each ranking, \mathbf{R}_1 and \mathbf{R}_2 .

The metrics to compare the rankings generated by our methodology with others in the literature are applied to evaluate the results found in the case study we present in Chapter 5. In this case study, we also evaluate the results when each measure of distance presented in Section 3.3.3 is applied to measure the cultural distance between countries to Brazil.

Finally, is important to point out that the methodology to measure cultural distance between countries, can also be used for other purposes besides a simple metric of distance. First, the value of cultural distance can be used as one more variable in gravity-type models to explain the flows of people between countries. For entertainment and marketing, for example, recommendation systems can be developed in order to recommend destinations for tourism based on the cultural similarities with a specific country. For economic reasons, the measure of cultural distance is important to define the most similar markets to make business. The methodology might be useful also for political and social applications, especially regarding diplomatic relations between countries, trade, and social problems related to immigrants.

In addition to the methodology to measure the cultural distance between countries, in the next section, we present the methodology to identify the cultural identity of countries. By this methodology, it is possible to characterize countries in terms of the cultural interests shared by the population.

3.4 Cultural identity

To answer the question **How to identify the cultural aspects that make up the identity of countries?**, we dedicate this section to model this specific problem given the data we have. The methodology for characterize the cultural identity of countries must contain:

1. The definition of the concept identity which will be presented in Section 3.4.1.

2. A description of the process of data normalization in order to create a probability distribution for each interest over the countries which will be described in Section 3.4.2.
3. Experiments using some different measures used to identify the cultural identity of a country which will be presented in Section 3.4.3.

3.4.1 Problem Definition

Identity: According to the dictionary definition, identity is the reputation or characteristics of a person or organization that makes the public think about them in a particular way⁹. In our context, we want to represent the cultural identity of a country in terms of interests related to cultural aspects. Informally, an interest is part of the cultural identity of a country if, when asked about where this interest comes from, one would answer only that country.

By our definition, identifying an interest as part of the cultural identity of a country corresponds to an interest that is associated with only one country. If one interest is associated with more than one country, it will not be considered as part of the cultural identity of any country. For example, if we think about some typical Brazilian dishes, like “Coxinha” and “Feijoada”, it is clear they are associated with Brazil. However, if we think about “Rice”, there is no association with only one country. Notice that an interest will not be part of the cultural identity of more than one country, however, a country could be characterized by more than one interest. In the example, Brazil is characterized by two interests, “Coxinha” and “Feijoada”.

The main goal by developing this new approach to characterize the cultural identity of a country is to identify interests that are associated with only one country. Due to the fact that we are considering a one-to-one relation, by identifying the country which is associated with the interest, we can also include the interest as part of the cultural identity of the country.

In Chapter 5, we present a case study in order to apply our methodology in the context of the interest around the world by typical Brazilian dishes. We identify the interests that are part of the Brazilian cultural identity given a subset of typical Brazilian dishes. The same methodology can be applied to identify the relationship between other subsets of interests and countries in order to characterize the countries in terms of their cultural identity.

⁹<https://dictionary.cambridge.org/us/dictionary/english/identity>

To model this problem, there are many different ways to identify the relationship between interests and countries. In the next sections, we will present the data and methods that can be used in order to characterize the cultural identity of countries.

3.4.2 Input: interest probability distribution

In order to find interests that are part of the cultural identity of a country, the first step is to define the data representation. After normalizing the data by the population as shown in Section 3.2, each cell $M'_{i,c}$ in matrix M' represents the proportion of the population in c who are interested in i . Now, we are interested in the association between interests and countries. The main idea is to find interests that can represent each country in terms of cultural identity. In order to find the interest that is part of the identity of each country, we would like to be able to compare the importance of each country for each interest. So, each interest is represented as a probability distribution over the countries. By doing this normalization for each interest, we construct another matrix M'' where each cell $M''_{i,c}$ is calculated as shown by Equation 3.11. Due to this, all the cells have a value between 0 (zero) and 1 (one) and the sum of values in each row is equal to 1.

$$M''_{i,c} = \frac{M'_{i,c}}{\sum_{c^*} M'_{i,c^*}} \quad (3.11)$$

The probability that we describe in Equation 3.11 is the probability of selecting a random person who is interested in i , and that person is from the country c . We will call this conditional probability: $Pr(c|i)$.

Therefore, by the conditional probability definition, we can define $Pr(c|i)$ as follows:

$$Pr(c|i) = \frac{Pr(c, i)}{Pr(i)},$$

where $Pr(c, i) = M'_{i,c} = \frac{A(i,c)}{A(c)}$ is the probability that the interest will be realized by a person from the country selected at random and $Pr(i) = \sum_{c^*} Pr(c^*, i)$ is the probability that the interest will be realized in the experiment.

In other words, after normalizing the audiences over the different population sizes, we are looking at the distribution of $P(c|i)$. To better explain the distribution we calculate by Equation 3.11, the experiment below describes exactly the meaning of the interest distribution. The interest distribution of an interest i corresponds to the i^{th} row in matrix M'' .

Experiment to explain the matrix M'' :

1. Initially, we will select k (e.g. 100) people from each country c .
2. For each pair, (c, k) , there is a random variable $X_{i,c}$ for the number of people from the country c who are interested in i .
3. Considering the random variable, $X_{i,c}$, as a binomial with a probability of success $Pr(c, i)$ given by the proportion of people in that country c who are interested in i , so $X_{i,c} = k * Pr(c, i)$. Notice that the probability of success, $Pr(c, i)$, is exactly the cell $M'_{i,c}$ from the matrix M' , that corresponds to the audience data normalized by the population in each country.
4. Thus, we have for each country c the expected value $E[X_{i,c}]$, which is the expected number of people living in the country c who are interested in i .
5. Now, of all the $k * m$ selected people, where m corresponds to the number of countries, we will separate the people who are interested in i and select one of these people.
6. The question we are answering with this input is: What is the probability distribution for that person to be from each of the m countries?

Given the description of the data, we would like to define an automatic method to identify interests that express the identity of each country. Given a set of interests and a set of countries normalized by Equation 3.11, we would like to represent the countries in terms of the interests that better represent them.

In the next section, we present methods that can be used to identify the interests that are part of the cultural identity of a country. To exemplify each method, we create a normalized matrix M'' with dimension 7×10 , as shown by Figure 3.2. We have seven interests, each one with a probability distribution over 10 countries, as shown in Figure 3.1. The examples present some interest probability distributions we would like to identify the country which could be characterized by the interest. Except for the first interest, shown in Figure 3.1a, which has a uniform distribution over the countries, for all the interests, the first country has a higher probability in comparison with the others. However, the distributions are different from each other.

Notice that, even with a small probability, for all the examples we assume a probability distribution with probabilities different from zero. Considering the Interest 8, shown in Figure 3.1h, for example, the probability of the first and second country is respectively, 0.95 and 0.045, the remaining error, $\epsilon = 0.005$ is uniformly distributed over the 8 countries. This assumption is important, mainly for the entropy difference measure, presented in Section 3.4.3. The entropy difference is based on a measure of

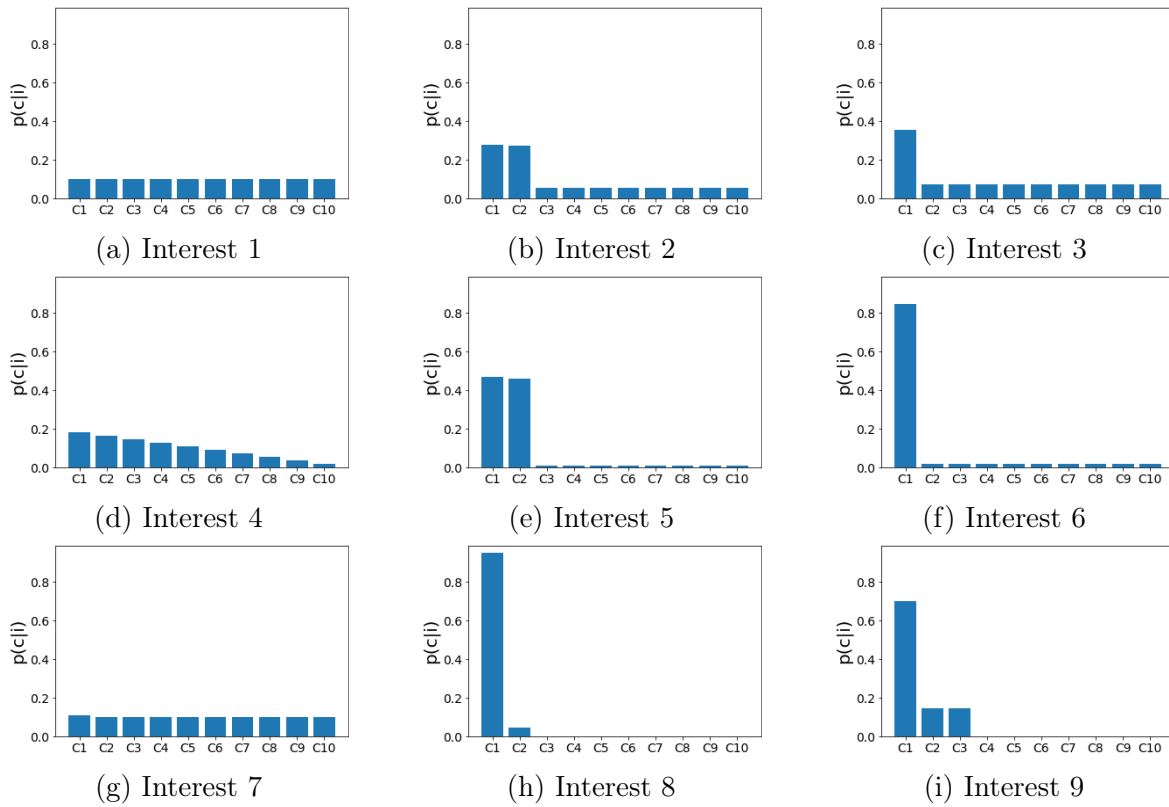


Figure 3.1: Examples of interests probabilities distributions.

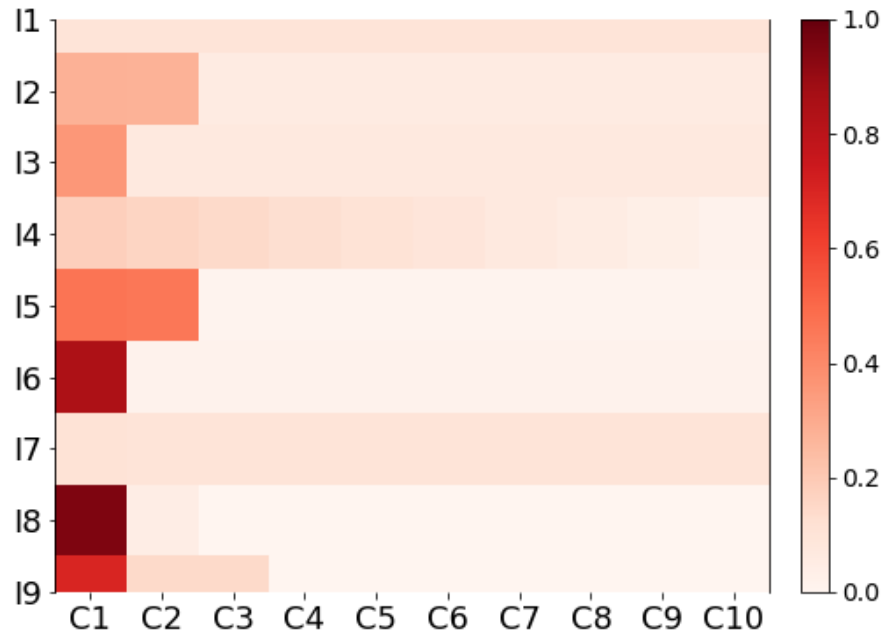
entropy, which is dependent on the distribution over the countries and the number of countries considered, when some probabilities are equal to zero, the measure tends to be lower than when all the countries have at least an infinitesimal probability. By adding a small probability uniformly distributed over the countries originally with a probability equal to zero, we want to reduce the negative impact on the result caused by these probabilities.

All the interest probability distributions in Figure 3.1 are used to exemplify the measures of cultural identity presented in the next section. In the next section, we also discuss the properties of each one of the measures, including the measure we propose, named entropy difference.

3.4.3 Measures of identity

There are many different measures that could be applied in order to solve the problem described in Section 3.4.1. Basically, the measures of identity are represented by an identity function \mathcal{I} as follows:

$$\mathcal{I} : \mathbf{I} \rightarrow \mathbf{C} \cup \{\emptyset\}$$

Figure 3.2: Matrix M'' .

Given a set of interests, each one of them will be associated with a country, in case they are part of the cultural identity of the country. Because of this, given a set of interests the function will return a set of countries or an empty set when the interests are not part of the cultural identity of any country.

In this section, we will present three possible measures, including the entropy difference, the measure we propose to identify interests that are part of the cultural identity of countries. We also compare the results obtained by each measure when applied to a bunch of examples in order to demonstrate the properties of the measure we developed. The entropy difference is also applied in the case study we present in Chapter 5. For all definitions, we assume that each interest i is represented by a probability distribution $Pr(c|i)$ over the countries c .

argmax:

The argmax method associates the interest to the country that has the highest probability. In other words, this measure identifies the interest as part of the cultural identity of the country if the conditional probability $Pr(c|i)$ is the highest in that country which dominates the interest in i . Given a probability distribution over all the countries c given an interest i , $Pr(c|i)$, the interest will be the identity of the country which has

Metric	I1	I2	I3	I4	I5	I6	I7	I8	I9
Argmax	✗	✓	✓	✓	✓	✓	✓	✓	✓
Entropy difference	✗	✗	✓	✗	✗	✓	✗	✓	✗
Inverse Simpson index	✗	✗	✗	✗	✗	✓	✗	✓	✗

Table 3.4: Interest identified as part of the cultural identity of a country considering each example of interest probability distributions for each measure of cultural identity. The green markers represent interests that are part of the cultural identity of a country while the red markers represent the interests that could not be associated with the cultural identity of only one country.

the highest probability as shown in Equation 3.12.

$$\mathcal{I}(i) = \operatorname{argmax}_c \{M''_{i,c}\} \quad (3.12)$$

The main issue with this definition is the fact that we are not considering the distribution as a whole and, because of this, we do not have a threshold to say how big the probability should be in comparison with other countries. Considering the examples in Figure 3.1 if we apply the Equation 3.12, except for the first interest which has a uniform probability distribution, the first country will be characterized by all the interests, as shown in Table 3.4. But when we look at some distributions, such as the probability distribution of Interest 7, it is clear that the first country does not dominate the distribution which is almost uniform. In this case, even if we observe the whole distribution, it is difficult to define a threshold to consider if the interest is part of the cultural identity of a country or not. Because of this, we want to develop a methodology to decide in an automatic way if there exists relationship between an interest and a country or not.

Entropy difference:

In this section, we will present the methodology to identify interests that are part of the cultural identity of countries. This methodology is based on an information theoretic approach that considers spatial analysis [Brodersen et al., 2012] to infer cultural identity.

Interest entropy and Interest focus give us an idea of how the interests are distributed around the globe. Given that $M''_{i,c}$ is the normalized proportion of the audience of a country c interested in i , the metric interest focus is given by Equation 3.13. This measure describes the proportion of the audience that is interested in i in a specific location. On the other hand, Equation 3.14 shows the interest entropy formula that

corresponds to the entropy measure of an interest distributed over the countries.

$$F_{i,c} = \frac{M''_{i,c}}{\sum_{c^*} M''_{i,c^*}} \quad (3.13)$$

$$H_i = - \sum_{c^*} F_{i,c^*} \log_2 F_{i,c^*} \quad (3.14)$$

These metrics provide support for evaluating the spread of various interests around the globe. Interests that are very concentrated in a region tend to manifest a low interest entropy and a high interest focus considering that region. It is particularly enticing to look at the interests that have moderate interest entropy because these interests are not entirely local, while also not completely common around the world. These interests could originate from a specific region and spread in popularity across other regions that share different cultural aspects, or become popular due to migration.

In order to find the relationship between interests and countries, for each interest, we will measure the entropy of its vector normalized over the countries. The entropy will measure exactly the uncertainty about the country in which the interest is part of the cultural identity. For a uniform distribution as shown in Figure 3.1a, for example, the uncertainty is maximum. If we consider extreme cases, like the Interest 6 shown in Figure 3.1f, the entropy is low due to the fact that the highest probability that corresponds to more than 0.8 is associated with one country. The point is we want to classify this relationship between interests and countries automatically. More important, we do not want to define a threshold to classify the value of entropy that will be associated with an interest that is part of the cultural identity of a country.

In order to classify these distributions in an automatic way, we will model the problem by using the Information Theory concept of Information Leakage. Considering the distribution regarding a specific interest i normalized over the countries, we can define this distribution as a prior distribution, π_i . We can see this distribution as previous knowledge. We now wish to quantify the information leakage caused by a channel C that processes the distribution π_i . Notice that we refer to a deterministic channel, where each input π_i leads to a unique output value H_i which we describe as $C(\pi_i) = \pi_i^*$. Basically, the prior distribution π_i is given as an input for the channel C that generated the posterior distribution π_i^* . The channel will change the distribution by modifying the probability to zero in the country considered the candidate to have the interest as part of its cultural identity and re-normalizing the distribution. At the end of this process, we are interested in comparing the entropy of the distribution to measure the information leakage.

Notice that the key point to the entropy difference identifies an interest as part of the cultural identity of a country is how the probability of the country, p^* , influences the prior distribution, π_i . In other words, measure the information leakage caused by a channel C which changes the probability p^* to zero and gives as output the new distribution π_i^* .

The information leakage will be given by the difference of the entropy of the prior distribution H_{π_i} and the entropy of the posterior distribution $H_{\pi_i^*}$. For each interest, the difference between the interest entropy considering and not considering the candidate country in the vector of countries expresses changes related to the uncertainty of which country is the interest associated with. When we consider the country, the entropy tends to be lower if the interest is more popular there than in other countries. Because of this, we define that an interest is part of the cultural identity of a country when the difference $H_{\pi_i} - H_{\pi_i^*}$ by considering the country is lower than zero.

In order to solve the inequality, we can write the entropy measure of the prior distribution π_i and the posterior distribution π_i^* as follows:

$$H_{\pi_i} = - \left(\sum_{k=1}^{M-1} p'_k \log(p'_k) + p^* \log(p^*) \right)$$

$$H_{\pi_i^*} = - \sum_{k=1}^{M-1} p_k \log(p_k)$$

where p^* corresponds to the probability of the interest in the candidate country. The probabilities p_k and p'_k are not the same since they came from different probability distributions.

Notice that H_{π_i} corresponds to the entropy measure by considering the prior distribution over all the M countries. However, $H_{\pi_i^*}$ corresponds to the entropy measure by considering the posterior distribution over $M - 1$ countries, since the probability of the candidate country goes to zero after going through the channel. By updating the entropy measure of the prior distribution, H_{π_i} can be written as follows:

$$H_{\pi_i} = - \left(\sum_{k=1}^{M-1} p'_k \log(p'_k) + p^* \log(p^*) \right)$$

$$= - \left(\sum_{k=1}^{M-1} (1 - p^*) p_k \log[(1 - p^*) p_k] + p^* \log(p^*) \right)$$

$$\begin{aligned}
&= -(1-p^*) \sum_{k=1}^{M-1} p_k \log((1-p^*)p_k) - p^* \log(p^*) \\
&= -(1-p^*) \sum_{k=1}^{M-1} p_k (\log(p_k) + \log(1-p^*)) - p^* \log(p^*) \\
&= (p^* - 1) \sum_{k=1}^{M-1} (p_k \log(p_k) + p_k \log(1-p^*)) - p^* \log(p^*) \\
&= (p^* - 1) \left(\sum_{k=1}^{M-1} p_k \log(p_k) + \sum_{k=1}^{M-1} p_k \log(1-p^*) \right) - p^* \log(p^*) \\
&= (p^* - 1) \left(-H_{\pi_i^*} + \log(1-p^*) \sum_{k=1}^{M-1} p_k \right) - p^* \log(p^*) \\
&= (p^* - 1)(-H_{\pi_i^*} + \log(1-p^*)) - p^* \log(p^*) \\
&= -(p^* - 1)H_{\pi_i^*} + (p^* - 1)(\log(1-p^*)) - p^* \log(p^*)
\end{aligned}$$

We are interested in solving the inequality $H_{\pi_i} - H_{\pi_i^*} < 0$. By using the definitions of the entropy measures H_{π_i} and $H_{\pi_i^*}$ we find the relationship between the entropy measure of the posterior distribution $H_{\pi_i^*}$ and the probability of the interest in the candidate country p^* as follows:

$$\begin{aligned}
&H_{\pi_i} - H_{\pi_i^*} < 0 \\
&= -(p^* - 1)H_{\pi_i^*} + (p^* - 1)(\log(1-p^*)) - p^* \log(p^*) - H_{\pi_i^*} < 0 \\
&= -p^* \log(p^*) - (1-p^*) \log(1-p^*) < H_{\pi_i^*} (1+p^* - 1) \\
&= -p^* \log(p^*) - (1-p^*) \log(1-p^*) < H_{\pi_i^*} p^* \\
&= \frac{-p^* \log(p^*) - (1-p^*) \log(1-p^*)}{p^*} < H_{\pi_i^*}
\end{aligned}$$

By using the entropy difference measure, we are especially interested in identifying the threshold to consider an interest as part of the cultural identity of a country. The entropy difference measure depends not only on the probability of the candidate country but also depends on the distribution over the countries.

Figure 3.3 represents graphically the result found that correlates the entropy measure of the posterior distribution $H_{\pi_i^*}$ and the probability of the interest in the

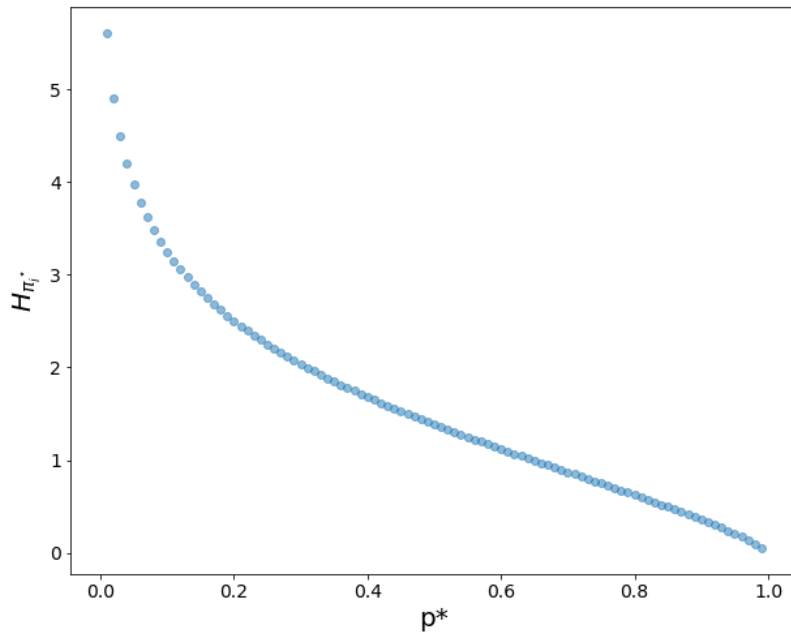


Figure 3.3: Relationship between the entropy measure of the posterior distribution $H_{\pi_i^*}$ and the probability of the interest in the candidate country p^* . The x-axis represents the values for the probability p^* and the blue curve represents the correspondent value for $H_{\pi_i^*}$, represented by the y-axis, following the equation $H_{\pi_i^*} = \frac{-p^* \log(p^*) - (1-p^*) \log(1-p^*)}{p^*}$.

candidate country p^* . The x-axis represents the values for the probability p^* and the blue curve represents the correspondent value for $H_{\pi_i^*}$, represented by the y-axis, given the probability p^* . We can interpret the graph as a threshold to identify the values for which an interest is considered part of the cultural identity of a country since points above the curve will represent exactly them.

By considering the examples in Figure 3.1, Table 3.4 shows the interest probability distributions for which the entropy difference measure identifies an interest as part of the cultural identity of a country. According to the entropy difference measure, interests 3, 6, and 8 are part of the cultural identity of the first country. For these three interests, the first country has a high probability in comparison with the other countries. However, for Interest 9, even the first country dominating the probability distribution, the interest is not considered part of the cultural identity of the country.

Notice that, especially for Interest 8, it is important to guarantee that there will not exist probability equal to zero. If we have the probability distribution concentrated only in two countries, the entropy measure of the posterior distribution, $H_{\pi_i^*}$, with only one probably country would be equal to zero. Given the inequality, $H_{\pi_i} - H_{\pi_i^*} < 0$, the

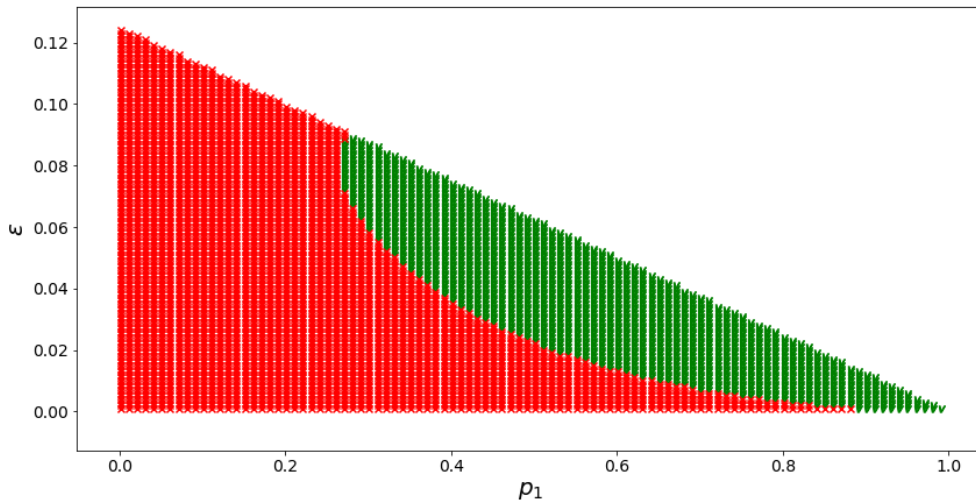


Figure 3.4: Example of identification when an interest is part of the cultural identity of a country by varying the probability of one country, p_1 , and ϵ . We are considering a distribution over two countries where $p_2 = 1 - (p_1 + \epsilon)$. The green and the red markers represent when the interest is or is not part of the cultural identity of one of those countries, respectively.

entropy measure of the prior distribution should be less than zero, which is impossible given the definition of entropy in Equation 3.14. It is important to have a probability distribution with no probabilities equal to zero before applying the entropy difference measure, a residual value, ϵ , is uniformly distributed over the countries with probability equal to zero in order to solve the problem.

To evaluate the impact of the probability distributions on the result, a set of distributions are created in order to identify the limits that make an interest be part of the cultural identity of a country by looking at the interest distribution over the countries. We are especially interested in examples such as interests 3, 5, 6, 8, and 9, where the probability is more concentrated in one or two countries. In the experiment, for all distributions, we vary the probability of the first country, p_1 , the second country, p_2 , and the residual value, ϵ , which is uniformly distributed over the $k - 2$ remaining countries. Notice that the probability of the second country is given by $1 - (p_1 + \epsilon)$. In this experiment, k is equal to 10.

Figure 3.4 shows the results where the x-axis represents the probability of the first country, p_1 , and the y-axis represents the residual value ϵ which corresponds to the sum of the probabilities of the $k - 2$ countries. The green points represent that the interest is part of the cultural identity of one of the two countries according to the entropy difference measure. Otherwise, a red point is used to mark. Notice

that Figure 3.4 shows exactly a threshold for which values of, p_1 and ϵ the interest probability distribution the entropy difference measure will be able to identify interests that are part of the cultural identity of a country.

Inverse Simpson index:

The Simpson index [Simpson, 1949], measures the degree of concentration when individuals are classified into types. This measure represents the probability that two entities taken at random represent the same type. However, we are particularly interested in the Inverse Simpson index, an index used to measure the Effective Number of Parties [Laakso and Taagepera, 1979].

Conceptually, the Effective Number of Parties is simply the number of “viable” or “important” political parties in a party system that includes parties of unequal sizes. The number of parties usually determines the number of effective parties, or how fragmented a party system is. Inspired by this theory, our goal is to use this measure to have an idea of the number of “important” countries for a specific interest. The Inverse Simpson index is computed by the formula shown in Equation 3.15.

$$N_i = \frac{1}{\sum_{c^*} M''_{i,c^*}{}^2} \quad (3.15)$$

where $M''_{i,c^*}{}^2$ corresponds to the square of each country’s proportion interested in interest.

By using the Inverse Simpson index, we identify the number of countries that are more “important” or, in other words, contribute more with the higher probabilities. If the Inverse Simpson index for an interest probability distribution is equal to 1, we conclude that only one country is important for that interest and then we can use the argmax measure, defined by Equation 3.12, to identify the country. In case of an Inverse Simpson index bigger than 1, the interest will not be consider as part of the cultural identity of the country. Equation 3.16 summarizes this explanation.

$$\mathcal{I}(i) = \begin{cases} \operatorname{argmax}_c \{M''_{i,c}\}, & N_i = 1; \\ \emptyset, & N_i > 1 \end{cases} \quad (3.16)$$

We can see that for completely unbalanced distributions with a dominant country, as for interests 6 (Figura 3.1f) and 8 (Figura 3.1h), the Inverse Simpson index is 1. It means that the first country dominates the distribution. However, even in unbalanced distributions with only one dominant country, like Interest 3 (Figure 3.1c), the Inverse Simpson index is higher than one because of the small probability of the

dominant country. It is clear that the first country dominates the distribution and, if the first country is removed, the remaining distribution is uniform. Clearly, in a uniform distribution, and considering our definition of identity, the interest is not part of the cultural identity of a country. Table 3.4 shows the cases where the Inverse Simpson index does not identify the interests as part of the cultural identity of a country.

The measures presented in this section, especially the measure proposed, seems to be reliable to identify interests that are part of the cultural identity of a country. In other words, the measures are reasonable to characterize countries in terms of interests representing cultural attributes. In the case study, presented in Chapter 5, we evaluate the results of the measure entropy difference in order to identify the Brazilian cultural identity in terms of typical Brazilian dishes.

As well as the methodology for measuring cultural distances, the methodology to identify interests that are part of the cultural identity of countries is important to the process of characterizing a country in terms of cultural aspects. The implications of this characterization are numerous, including the effect in businesses, marketing, migration, and diplomatic relations between countries.

In addition to the methodologies, in the next chapter, we propose the use of online social media data to represent the audience interested in some interests for each country. In Chapter 4 we present the Facebook Advertising Platform with examples of how to collect the data. As an example of usage, we present the data collected regarding typical Brazilian dishes, used in the case study in Chapter 5.

Chapter 4

Gathering Facebook Data

In chapter 3 we present our methodology to characterize the cultural identity of countries in terms of interests related to cultural attributes. We also present our methodology to measure the cultural distance between countries. Both methodologies include the problem description, input description, and measures that can be applied in order to solve the problems. In this chapter, we present our data collection strategy as a way to enable a Case Study presented in Chapter 5 in which we use our methodology to characterize the Brazilian cultural identity in terms of typical Brazilian dishes.

4.1 Background on Facebook Ads

With almost 2.5 billion monthly active users as of the fourth quarter of 2019, Facebook is the most popular online social network¹. In addition to the most famous social network, Facebook is a social network with very active users. A recent survey [Smith et al., 2018] indicates that two-thirds of US adults use social networks, out of which 42% say it would be hard to give up social media. The same report shows that 51% of US Facebook users access their accounts several times per day.

Due to a large number of users, the advertising tools provided by Facebook and other social networks, in general, have been widely explored by companies looking for their target audience on social networks. One of the keys to the success of online social network advertising platforms is the vast possibilities to reach users such as by providing a list of personally identifiable information (name, phone number, email, etc) or by configuring targeting options from a huge list of fine-grained attributes such as race, income level, interests, and behaviors.

¹<https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>

In this work, we are interested in characterize countries in terms of cultural identity and also measure the cultural distance between countries. In order to collect representative data regarding cultural interests, we explore the simulation of advertisements in order to collect the Facebook user’s audience regarding some interests representing cultural aspects around the world.

4.2 Collecting Facebook Ads data

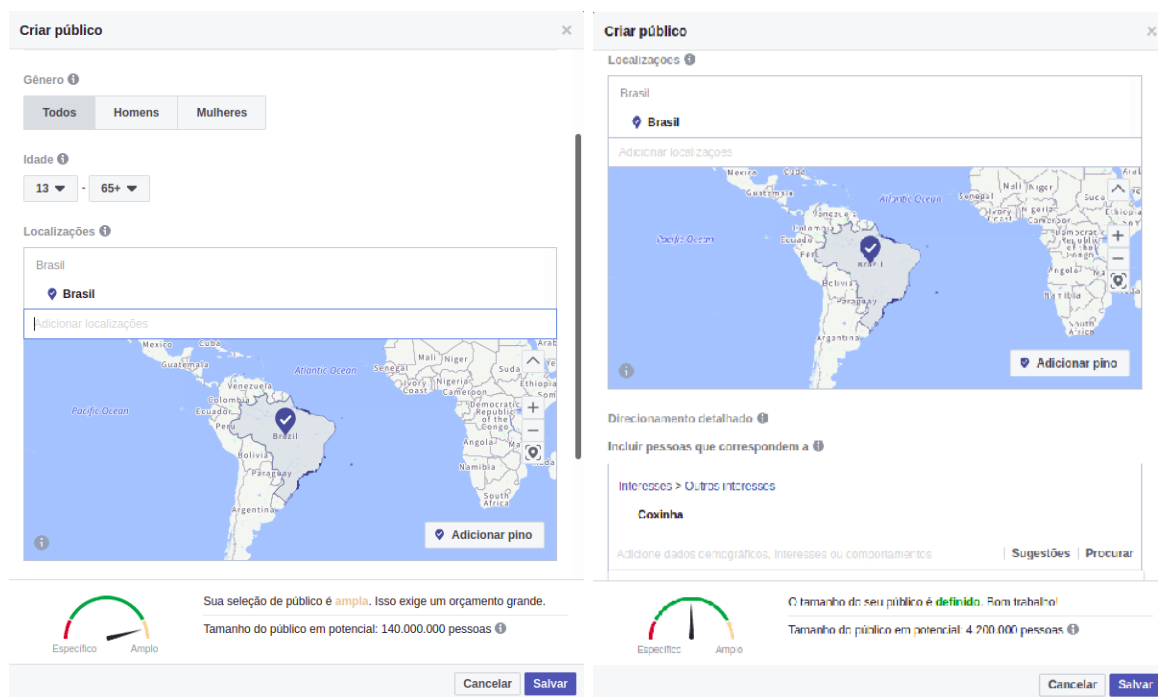
In this section, we explain how the simulation of advertisements works in order to collect the Facebook users audience regarding some interests around the world.

The platform to create an advertisement on Facebook, namely Facebook Ads, allows users to compute an estimated audience size for a proposed advertisement [Kosinski et al., 2015]. This audience can be defined by demographic attributes provided by Facebook, including gender, age, home location, and interests, which can be informed by the user or inferred by Facebook based on user’s likes or status updates. On Facebook “pages” and “interests” are two completely different things, but they can be associated. There are millions of Facebook pages, but not all of these are associated with an interest that can be targeted with ads.

Basically, Facebook users generate traces of their preferences over multiple domains such as music, food, and, sports [Dubois et al., 2018]. In this thesis, the Case Study presented in Chapter 5 focuses on typical Brazilian dishes as a marker of cultural distance between countries.

As an example of attribute targeting, one may select users who live in a city (e.g. Belo Horizonte) or country (e.g. Brazil) that are interested in one typical dish (e.g. Feijoada, Coxinha). In this example of usage, the owner of a restaurant might be interested in selling Coxinha at the restaurant and spread the news by showing ads to Facebook users who are interested in this kind of food and live in a particular city. Basically, the main usage of these tools is to advertise a product or service by selecting the target audience. However, an important application can simulate an advertisement to know the audience of an specific interest before investing in that field, for example.

Figure 4.1 depicts some examples of a target audience that may be constructed using some of the attributes provided by the Facebook Advertising Platform. At first, we defined the location as Brazil (see figure 4.1a). The potential target for this particular combination of attributes is shown by Facebook before running the ad, in this case, 140.000.000 users. This number represents the maximum number of users an ad with this target options might reach. The actual number of users the ad will reach



(a) Targeting Facebook users who lives in Brazil (b) Targeting Facebook users interested in Coxinha who lives in Brazil

Figure 4.1: Facebook Ads Platform.

after its publication is not necessarily that potential target, but rather depends on how much money the advertiser decides to pay.

Notice that on the top of the location selection, we can also choose the range of the age we would like to consider and the gender. To collect the audience considering the whole population in Brazil, we select both gender and the minimum and maximum limit of age between 13 and 65 or more. Next, as shown in figure 4.1b we define our target as all Facebook users living in Brazil who are interested in Coxinha, a typical dish in Brazil. The audience for this targeting specification is 4.200.000 (3% of the Facebook users living in Brazil).

By considering the case study presented the Facebook Ads API² available for Python³ is used to collect the audience of typical Brazilian dishes on Facebook based on their users' registered preferences. This tool serves our purposes well because of this the audience can also be collected within a given country, as most users have their home location registered in the system. In Chapter 5, Figure 5.1 shows the proportion of Facebook users living in a country in comparison with the real population.

As an example, Table 4.1 shows the audience for some typical Brazilian dishes

²<https://developers.facebook.com/docs/marketing-apis/>

³<https://pypi.org/project/facebookads/>

Interest audience	
Açaí	16168360
Coxinha	6305820
Mandioca	14658390
Misto-quente	631070
Pão de queijo	441630
Rabanada	2784250
Requeijão	215680
Tapioca	5910930
Torresmo	2758570
Sobá	1518770
Vagem	5047300
Chouriço	8945210
Churrasco	350213590
Cuscuz	4598380
Feijoada	3900700
Arroz	142233190
Feijão	6285700
Estrogonofe	1090980
Cachaça	5413560
Caipirinha	3532140

Table 4.1: Facebook audience for some typical Brazilian dishes.

by considering the Facebook user’s preferences around the world. The data collected will be better described in the case study presented in Chapter 5.

Although the large amount of data available in the Facebook Advertising Platform, it is important to disclose the API limitations as well as the steps taken to preserve the privacy of users. In the next section, we present some API limitations and discuss the privacy of Facebook user’s data.

4.3 Privacy and Limitations

First of all, it is important to point out that, despite the several privacy issues that surround the online social network advertising platform, our methodology does not represent a threat to users’ privacy. One may argue that our methodology may lead to the leakage of personal information about users, however, our work uses only aggregate information. It is only a number that represents the quantity of users that match each particular set of attributes. In addition to this, there is no need to run any ad in our approach, meaning that all gathered data is returned before any cost is incurred. Therefore, we do not collect any personally identifiable information. Among

the millions of requests we performed to the Marketing API, none of them were able to gather and link any personal information to any particular user.

If the audience size is between 0 (zero) and 1000 (one thousand), the Facebook Ads API will return the default value of 1000. Because of this restriction, the collection considering specific interests inside a small population may not give information about the exact number of Facebook users that match the criteria specified. Therefore, the comparison in terms of interest in typical Brazilian dishes between Brazilian expats and the rest of the population, especially in countries with a small audience on Facebook, is not reliable. Thus, for all cases the API returned the default value, we set the audience to 0.

Though the Facebook Advertising Platform can be explored to infer demographics from the offline world, the mechanisms behind the tool are not publicly known, which is a limitation of our method. As a black box, make it difficult for the researchers to check if and to what extent the data is reliable. Furthermore, the population of Facebook is known to be biased towards gender, age, and other aspects as previously discussed by some authors [Araujo et al., 2017]. Therefore, it is important that conducted studies relying on this methodology validate their data. On the other hand, these issues open new research avenues on the statistics and demographics that might apply their artifacts to deal with noise and imperfect data in order to improve the confidence of the data.

The lack of control in the attributes set is another limitation of this research. As stated before, missing attributes may occur because Facebook has no interest in creating or keeping it. This situation may produce some inconsistencies in the final dataset, since a very popular entity, such as a largely known newspaper may have no interest related to it, whereas a low popular media outlet has a related attribute. Furthermore, some attributes are summarily discontinued without explanation or previous warning, which means that studies aiming to explore the evolution of demographic audiences for certain entities or studies that evaluate the audience of ads run in the past may eventually face a lack of data.

Despite the limitations, all previous works, described in Section 2.3 provide evidence that Facebook Ads data can indeed be used as a source of information. In the next chapter, by using Facebook Ads data, we present the case study of how our methodology can be used to identify cultural aspects of the Brazilian cultural identity and how to measure the cultural distance of other countries to Brazil.

Chapter 5

The Case of Brazilian Cuisine

In this chapter, typical Brazilian dishes are used as a proxy of the Brazilian cultural identity by applying the methodology described in Section 3.4. From this, we also measure the cultural distance, as presented in Section 3.3, between Brazil and the countries with more Brazilian immigrants. To make these analyses feasible, we use the data collected from the Facebook Advertising Platform as described in Chapter 4.

The rest of this chapter is organized as follows. First, we describe the data collected. Then, we apply the methodology to measure the cultural distance, including the normalization process. Next, we detail the methodology to infer the Brazilian cultural identity. Finally, we discuss our findings regarding the results obtained by the methodologies.

5.1 Data description

Since there is a great variety of typical local food in Brazil, we selected a set of the 20 most popular Brazilian dishes according to BBC Good Food¹ and the list of Brazilian dishes available on Wikipedia². Due to the fact that our main goal is to compare various countries with Brazil in terms of cultural distance, for this comparison we selected a set containing the 29 countries most preferred by Brazilian immigrants according to the Ministry of Foreign Affairs, Itamaraty³. All the subsequent analyses focus on these sets of typical Brazilian dishes and countries and the data collected following the description given in Chapter 4.

¹<https://www.bbcgoodfood.com/howto/guide/top-10-foods-try-brazil>

²https://en.wikipedia.org/wiki/List_of_Brazilian_dishes

³<http://www.brasileirosnomundo.itamaraty.gov.br/a-comunidade/estimativas-populacionais-das-comunidades/Estimativas%20RCN%202015%20-%20Atualizado.pdf>

Figure 5.1 shows the number of users in each location as well as the actual population of the countries. Some countries, like China, will evidently not provide a good estimate for the actual population, given that the number of Chinese users on Facebook is under 0.1% of the real population. Nevertheless, in other countries more than 50% of the real population is part of the Facebook audience.

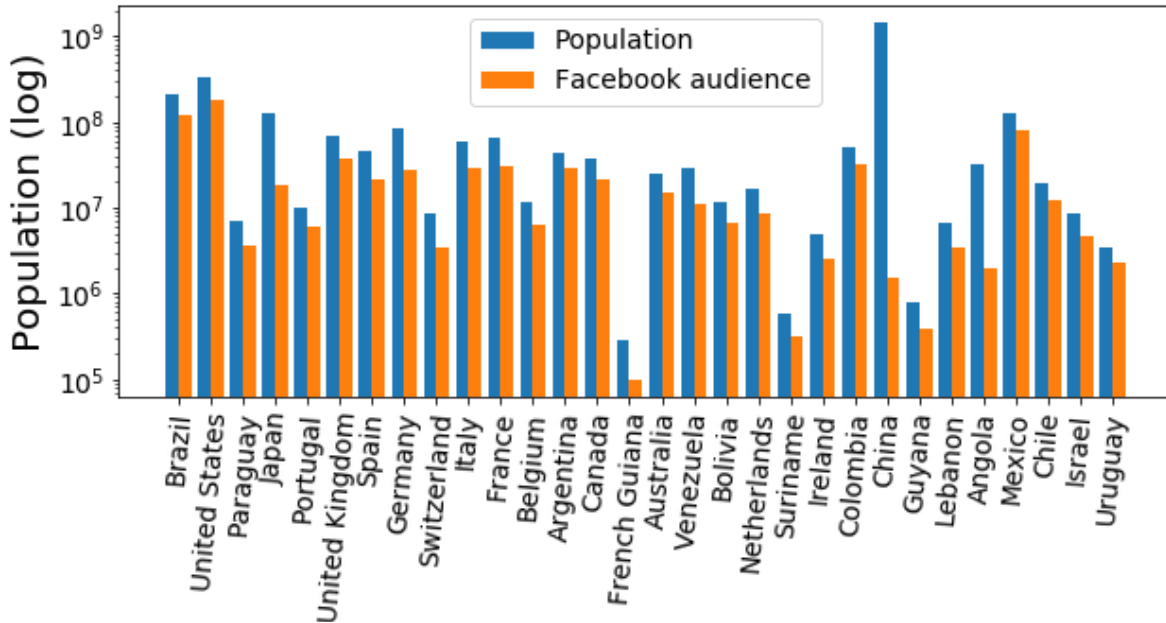


Figure 5.1: Real population and Facebook audience in each country (log scale).

As discussed in Section 3.2, Figure 5.1 also shows that the size of the audience in each country can vary a great deal across countries. Thus, to make a fair comparison between interests in these countries, we need to normalize the audience in each interest by the estimated Facebook population in each country by using the Equation 3.1 presented in Chapter 3. After normalizing the data, we are able to apply the methodologies to measure the cultural distance and to characterize the country in terms of interests related to cultural attributes, as described in the next sections.

5.2 Brazilian cultural distance

As a first step to calculate the cultural distance between countries, the countries should be represented by a vector. Figure 5.2a shows the matrix containing the normalized audience interested in Brazilian dishes for all countries. In this case, note that in all the selected countries, the highest interest is for “Churrasco” (“Barbecue”) and, in second, “Arroz” (“Rice”). If we consider each column as a vector representing the country interests, these unbalanced distributions bias the distance measurement between two

countries by these two most popular dishes. In order to give the same importance for all dishes, we normalize and smooth these distributions by their z-scores. We apply the Equation 3.2 shown in Chapter 3 for calculating the z-scores.

Figure 5.2b shows the heatmap after z-score normalization. As expected, we observe that the distribution is diverse and does not seem to exhibit a few dominant interests in all the countries.

The z-score normalization allows each country to be represented by a vector of preferences regarding typical Brazilian dishes. The aim is to compare those individual vectors with the benchmark Brazilian vector. The most similar countries to Brazil will exhibit small distances. After we generate a set of measures for each country, given by the distance between the country interest vectors and the Brazilian vector, we rank the countries according to the cultural proximity with Brazil. The ranking generated considering the cultural distance can be compared with other types of rankings that attempt to measure the similarity between countries, for instance, the ranking constructed with the most preferred countries by Brazilian immigrants, or Brazilian expats in each country according to Facebook, and rankings that express the geographic distance between countries. The rankings that will be used for this comparison are described next.

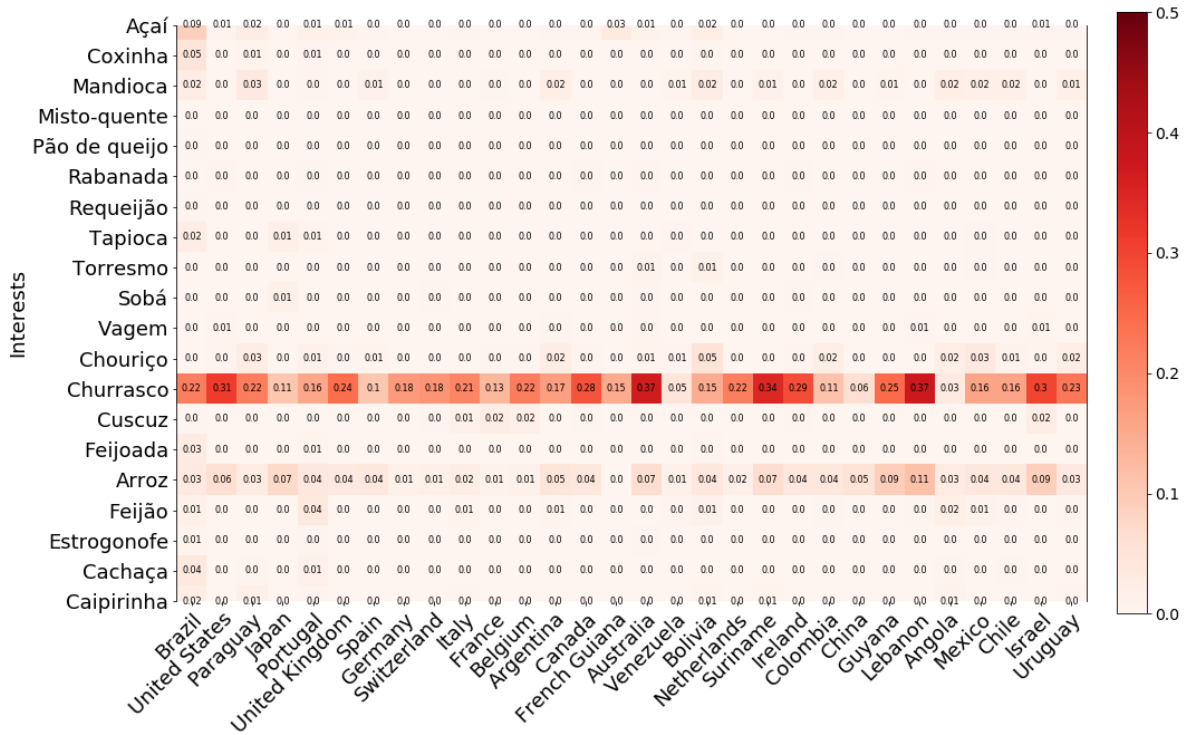
5.2.1 Baseline data

The ranking considering the cultural distance to Brazil can be constructed using different distance metrics of distance, and in this work we use: Euclidean, Cosine, Mean Absolute error, Relative error and Earth mover’s distance. For each metric, a ranking is generated and compared with the baselines below

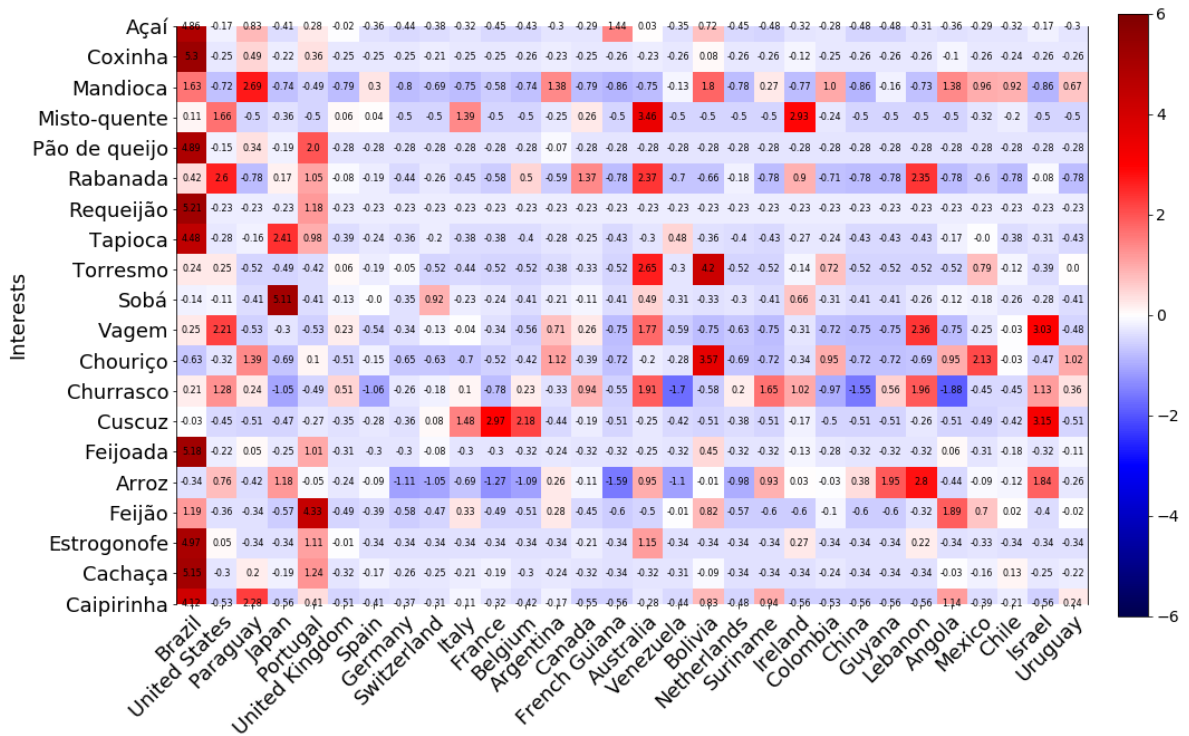
Immigrant ranking: Figure 5.3a shows the ranking of the countries that have more Brazilian immigrants in proportion to their real populations.

Expat (Facebook) ranking: Figure 5.3b shows the same countries presented in *Immigrant ranking* sorted by the countries that have more Brazilian expats in proportion to their audience according to Facebook Ads. We can see that both rankings, *Immigrant* and *Expat (Facebook)* are well correlated, while Facebook Ads seems to represent the proportion of Brazilians in those countries well.

Geographic distance ranking: The geographic distance can be expressed in terms of the simple geographic distance or in terms of the weighted distance [Mayer and Zignago, 2011]. The *Geographic distance ranking* is sorted by the countries that are most close to Brazil in terms of a simple geographic distance calculated following the great circle formula, which uses latitudes and longitudes of the

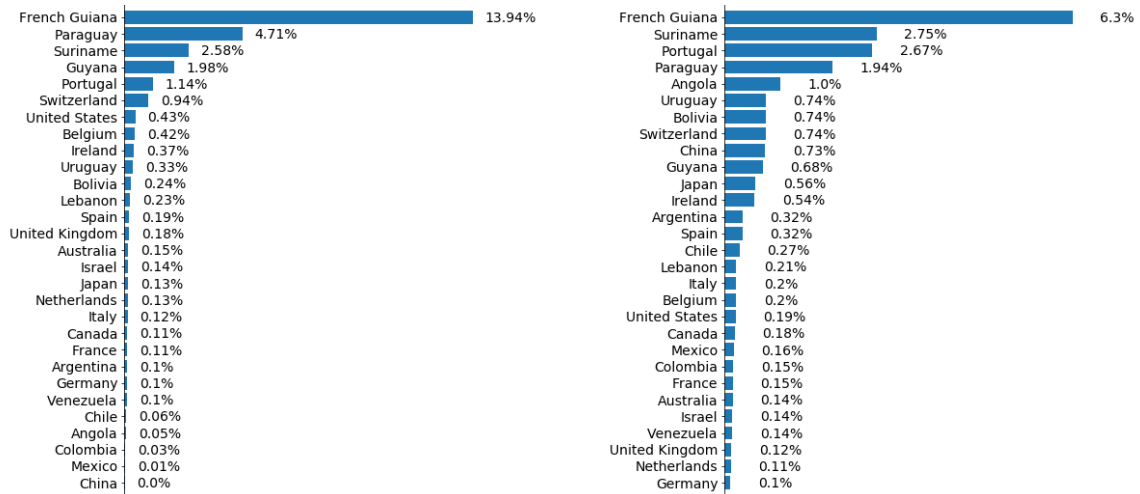


(a) Before z-score normalization.



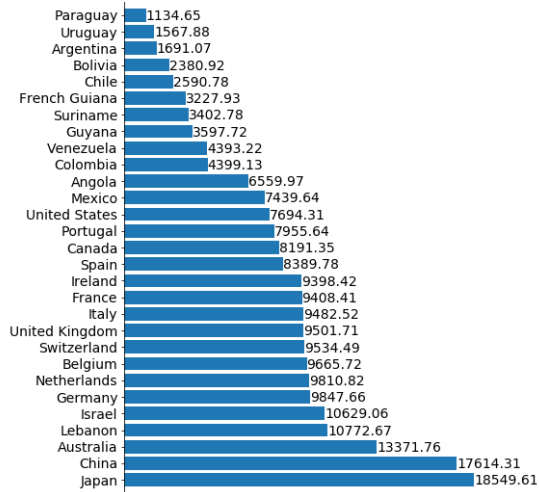
(b) After z-score normalization.

Figure 5.2: Proportion of interest in each country. All the interests are normalized by the audience in each country.

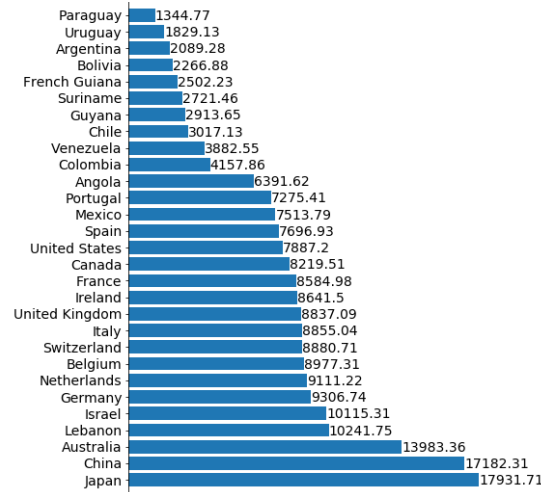


(a) *Immigrant ranking*: Proportion of Brazilian immigrants in real population.

(b) *Expat (Facebook) ranking*: Proportion of Brazilian expats in Facebook audience.



(c) *Geographic distance ranking*: Distance to Brazil.



(d) *Geographic weighted distance ranking*: Weighted distance to Brazil.

Figure 5.3: *Immigrant, Expat (Facebook) rankings, Geographic and Geographic weighted distance ranking.*

most important city in terms of population. *Geographic weighted distance ranking* also shows the countries that are most close to Brazil, considering the distance between the main agglomerations of all countries [Head and Mayer, 2002]. Since these two rankings are strongly correlated (0.96), Figure 5.3d shows only the *Geographic weighted distance ranking*.

Given the baseline rankings, we want to compare and measure the correlation between them and in comparison with the rankings created by the cultural distance between other countries to Brazil. In the next section we present the ranking correlations by applying the metrics described in Section 3.3.4.

5.2.2 Comparison between rankings

Given the rankings created by the cultural distance measure and the baseline rankings presented in Section 5.2.1, Table 5.1 shows correlations and p-values between the *Immigrant ranking*, *Expat (Facebook) ranking* and the rankings generated with different measures of distance. Also, the correlations between all of them and the *Geographic weighted distance ranking* are shown in Table 5.2. The metrics applied to calculate the correlations are: (i) **WT**: Weighted tau correlation; (ii) **KT**: Kendall tau correlation; (iii) **S**: Spearman correlation and (iv) **J**: Jaccard similarity considering the top 10 in each ranking.

Rankings	WT ⁽ⁱ⁾	KT ⁽ⁱⁱ⁾	S ⁽ⁱⁱⁱ⁾	J ^(iv)
Euclidean distance	0.2403	0.33 (0.01)	0.3818 (0.04)	0.3333
Cosine distance	0.3931	-0.0739 (0.57)	-0.0995 (0.61)	0.25
Mean Absolute Error	0.3396	0.0788 (0.55)	0.1197 (0.54)	0.1765
Relative Error	0.0099	0.0542 (0.68)	0.103 (0.60)	0.25
Eart Mover's distance	0.1072	0.1823 (0.17)	0.2596 (0.17)	0.25
<i>Expat (Facebook) ranking</i>	0.716	0.0296 (0.82)	0.0655 (0.74)	0.5385
<i>Geographic distance ranking</i>	0.452	-0.0296 (0.82)	-0.0443 (0.82)	0.3333
<i>Geographic W. distance ranking</i>	0.4675	0.0246 (0.85)	0.0128 (0.95)	0.3333

Table 5.1: Comparison with the *Immigrant ranking*.

Rankings	WT ⁽ⁱ⁾	KT ⁽ⁱⁱ⁾	S ⁽ⁱⁱⁱ⁾	J ^(iv)
Euclidean distance	0.1896	0.2611 (0.05)	0.3586 (0.06)	0.5385
Cosine distance	0.1833	0.1133 (0.39)	0.2015 (0.29)	0.1765
Mean Absolute Error	0.2181	-0.1675 (0.20)	-0.2517 (0.19)	0.25
Relative Error	-0.3877	0.0246 (0.85)	0.0236 (0.90)	0.25
Eart Mover's distance	-0.2908	0.0148 (0.91)	-0.001 (1.00)	0.1111
<i>Expat (Facebook) ranking</i>	0.4563	-0.0099 (0.94)	0.0172 (0.93)	0.4286
<i>Geographic distance ranking</i>	0.9678	0.6601 (0.00)	0.8108 (0.00)	1.0
<i>Immigrant ranking</i>	0.535	0.0246 (0.85)	0.0128 (0.95)	0.3333

Table 5.2: Comparison with the *Geographic weighted distance ranking*.

In addition to comparing the rankings, we are also interested in giving more importance to the first few countries because of their representation in terms of the fraction of Brazilian immigrants in the real population and in the Facebook audience. Hence, we decided to consider the measure of correlation that allots more weight to the top elements in the rank. The weight is mapped from non-negative integers (zero representing the most important element, the first in the *Immigrant ranking*) to a non-negative weight, given by a hyperbolic weighing. The hyperbolic weighting maps the

position of each element in rank r to a weight $\frac{1}{r+1}$. Because of this, the first element ($r = 0$) has a weight equal to 1, the second, $\frac{1}{2}$, and so on.

As shown in Table 5.1, considering the `Weighted tau` correlation, Cosine distance is the distance metric that generates the Cultural distance ranking that better approximates to the *Immigrant ranking* (almost 0.4 correlated). Figure 5.4 shows the comparison between them. Considering `Kendall tau` and `Spearmanr`, the correlation between Euclidean distance ranking and *Immigrant ranking* is higher when compared to other distances.

The Cosine distance ranking has a negative correlation when we consider `Kendall tau` and `Spearmanr`. This happens because the last countries in *Immigrant ranking* are associated with a non-matching position in the Cosine ranking list. But when the weights associated to all of the countries are listed in the decreasing order of importance in the *Immigrant ranking*, the mismatches that occur in those positions do not substantially impact the correlation.

The correlation between the *Immigrant ranking* and the *Expat (Facebook) ranking*, considering the `Weighted tau`, is more than 0.70. This high correlation shows that Facebook data can be a good estimator of Brazilian immigrants around the world. Also, the correlation between the *Immigrant ranking* and the *Geographic distance ranking* shows that migration is less correlated with the geographic distance, a correlation of 0.45. In fact, it is well known that there are other decisive factors that justify the migration, not only the proximity in terms of geographic distance [Faist, 2000].

Figure 5.4 shows exactly the correlation between migration and cultural distance. The *Immigrant ranking* corresponds to the percentage of Brazilian immigrants in some countries and the cultural distance is given by the cosine distance between countries to Brazil. The distance is calculated between vectors generated for each country considering the percentage of Facebook users who are interested in some typical Brazilian food. The most similar countries to Brazil are at the top of the second list in Figure 5.4. This figure should be interpreted as a comparison between rankings where the top 10 countries in the *Immigrant ranking* are shown in colors.

Comparing the *Immigrant ranking* and the Cosine distance ranking, shown in Figure 5.4, we see that despite the large proportion of Brazilian immigrants in countries like Switzerland (6th in *Immigrant ranking*), they seem not to be strongly attached to the Brazilian culture. The opposite is observed in countries like Portugal and Paraguay, that are most preferred by Brazilian immigrants and are most similar in terms of Brazilian food preferences to Brazil. Portugal is the country most similar to Brazil in terms of the preferences for the typical Brazilian dishes. This similarity cannot be related to the geographic distance since Portugal is not close to Brazil, but the

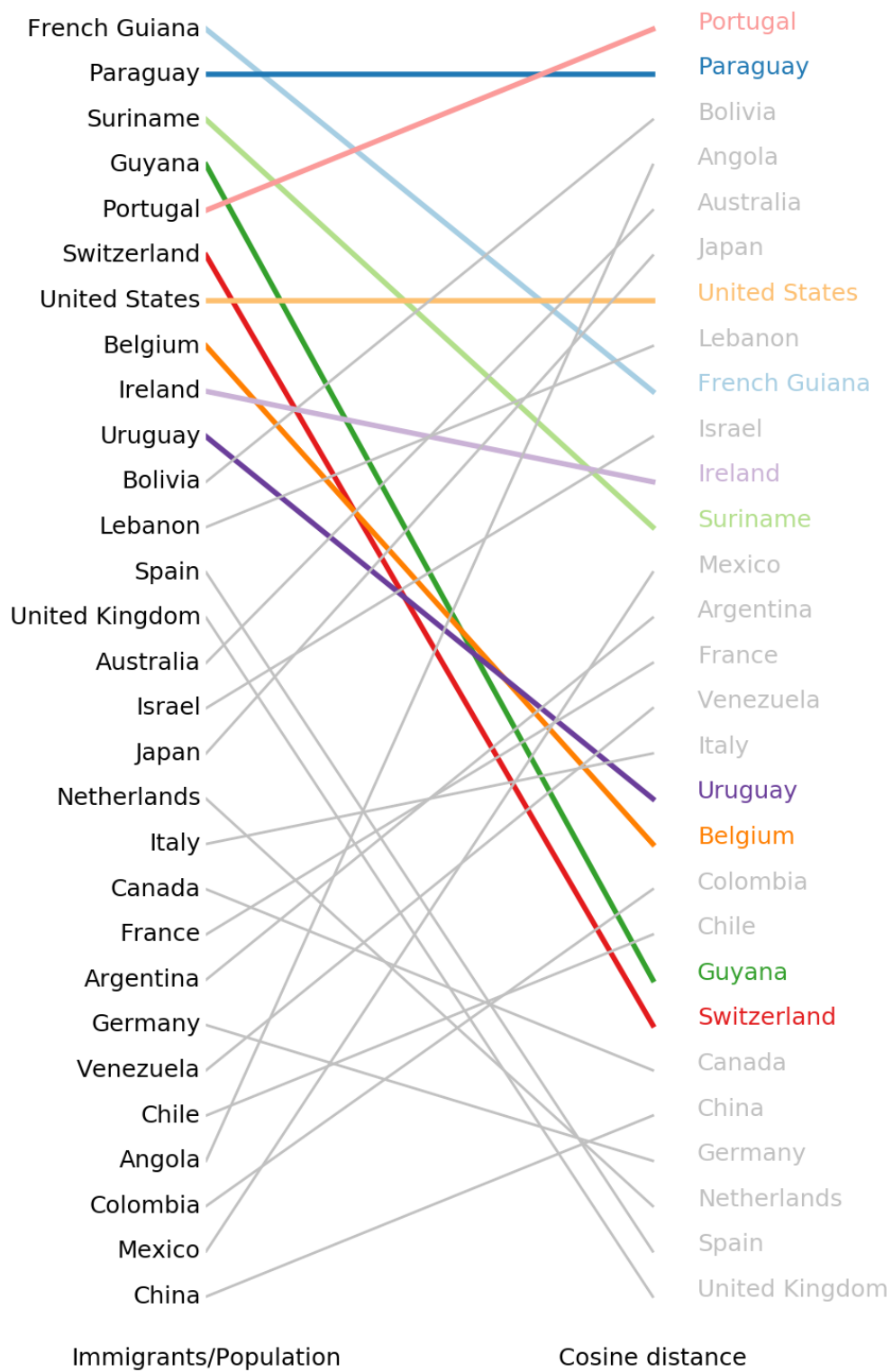


Figure 5.4: Comparison between *Immigrant* and Cosine distance rankings.

proportion of immigrants in Portugal is one of the highest according to the *Immigrant Ranking*. The language and the general cultural similarity [Kogut and Singh, 1988] shared between the former colony of Portugal, Brazil, explain in part the pull factors of migration to Portugal. Considering the United States, the country with more Brazilian immigrants in terms of absolute value, the position in the *Immigrant ranking* and in the Cosine distance ranking remains the same, so the similarity in terms of Brazilian food preferences are well correlated with the number of Brazilian immigrants in the population. Other countries, like Paraguay and Bolivia, seem to be more similar to Brazil in terms of food interests because of geographic proximity. In general, most of the countries like Argentina, Venezuela, Colombia, and Chile, that are geographically closer to Brazil are at a higher position relative to the cultural distance ranking, as compared to their *Immigrant ranking*. In this case, this result shows that the interests can be justified not only by the migration processes but also by geographic proximity, given that the local population seems to be interested in several typical dishes from the neighboring countries. However, generally, the migration process is one of the most crucial factors that expose distant countries (from Brazil) to Brazilian cuisine.

In general, the methodology proposed to measure the cultural distance, generates good results when applied to the case study of Brazilian cuisine. In addition to the methodology proposed to measure the cultural distance, in the next section, the methodology presented in Section 3.4 is applied to this case study in order to characterize Brazil in terms of its cultural identity.

5.3 Brazilian cultural identity

In the previous section, Figure 5.2b shows the proportion of Facebook users interested in some typical Brazilian dishes after the z-score normalization. In general, for most of the interests, Brazil is above average. “Chouriço”, “Arroz” (“Rice”, in English), “Sobá” and “Cuscuz” are the only interests below average, indicating that these interests may be more popular in other countries. If we consider that typical Brazilian foods are those in which the interest z-scores in Brazil are the highest among other countries, then 9 dishes meet this criterion: “Açaí”, “Coxinha”, “Pão de Queijo”, “Requeijão”, “Tapioca”, “Feijoada”, “Estrogonofe”, “Cachaça” and “Caipirinha”. Notice that, in this case, by selecting the interests which Brazil has the highest value is the same to select the interests which the argmax measure, defined by Equation 3.12, is Brazil. In this section, this approach will be compared to an information theoretic approach, presented in Section 3.4, which considers entropy analyses to identify the interests that are part

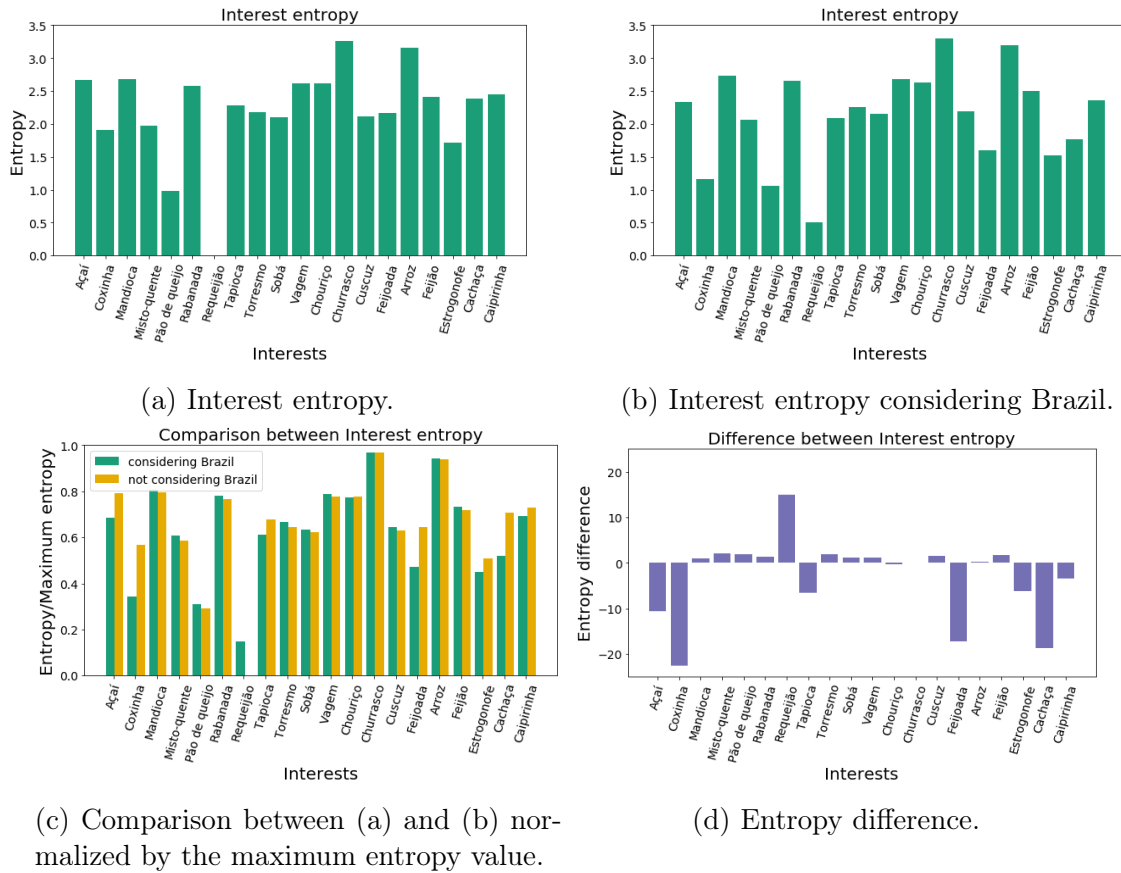


Figure 5.5: Comparison between the measure of interest entropy and entropy difference.

of the cultural identity of a specific country.

To understand the distribution of typical Brazilian interests around the world, we can consider the z-score normalization matrix and also make use of the methodology described in Section 3.4. The methodology, entropy difference, is applied to evaluate the distribution of interests regarding typical Brazilian dishes over countries to identify the interests that are part of the cultural identity of Brazil.

The entropy difference expresses changes related to the uncertainty of which country is the dish associated with. Notice that in this case study, Brazil is the candidate country. Because of this, when we consider Brazil in the prior probability distribution, the entropy is lower than the entropy of the posterior distribution, with the Brazilian probability equals to zero, if the interest is significantly more popular in Brazil than in other countries. Figure 5.5a and Figure 5.5b show, respectively, the value for the interest entropy measure for the posterior and prior distributions and Figure 5.5c shows the comparison between them. The results regarding the entropy difference measure are presented in Figure 5.5d. Notice that the seven interests identified as part of the Brazilian cultural identity are: “Açai”, “Coxinha”, “Tapioca”, “Feijoada”, “Estrogonofe”,

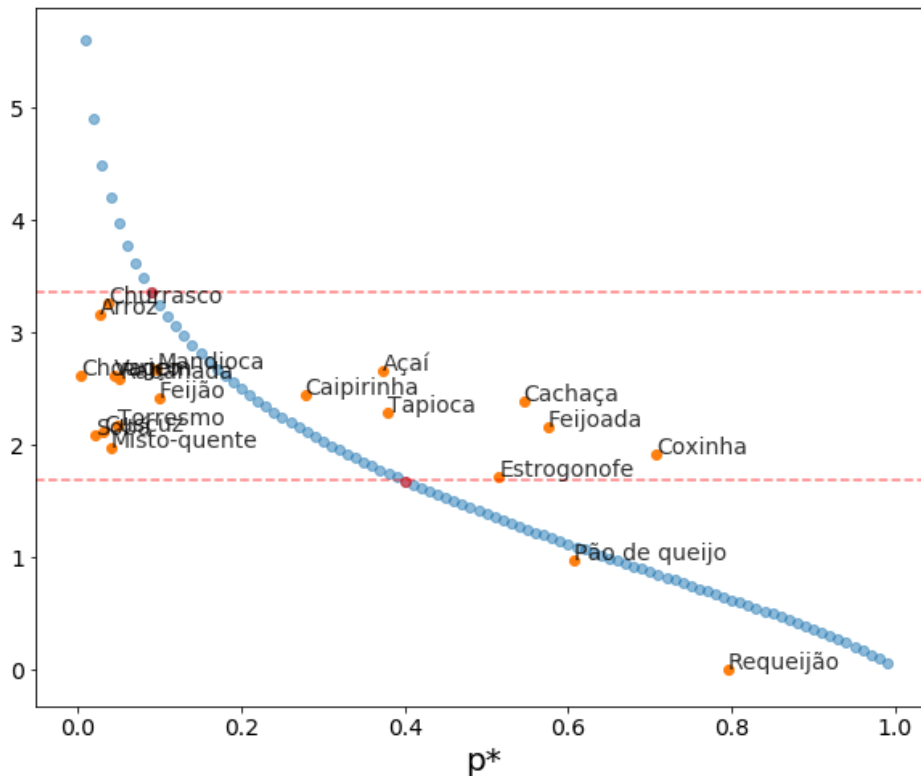


Figure 5.6: Relation between the entropy measure of the posterior distribution $H_{\pi_i^*}$ and the probability of the interest in Brazil p^* .

“Cachaça” and “Caipirinha”.

Figure 5.6 shows the same entropy difference curve in blue, shown in Figure 3.3. However, each point in orange corresponds to the interests which represent typical Brazilian dishes. The points above the curve correspond to interests with negative entropy differences. The red horizontal lines represent, respectively, the maximum and 50% of the maximum entropy of a vector with 29 positions. Notice that the probability distribution vector of the posterior distribution contains 29 positions corresponding to the 30 countries selected, except Brazil which probability goes to zero in the posterior probability distribution.

Finally, we compare the typical Brazilian interests according to these approaches with the 6 typical Brazilian dishes listed by BBC Good Food. Considering the argmax of the z-score normalization matrix and the interest entropy difference for each interest, we identify 5 and 4 common interests, respectively. The other 2 dishes from BBC Good Food that are not considered by our metrics as typical from Brazil, “Churrasco” and “Pão de Queijo”, seems to be popular in other countries as shown by the metric interest focus in Figure 5.7. We see that “Churrasco” has a uniform distribution over the

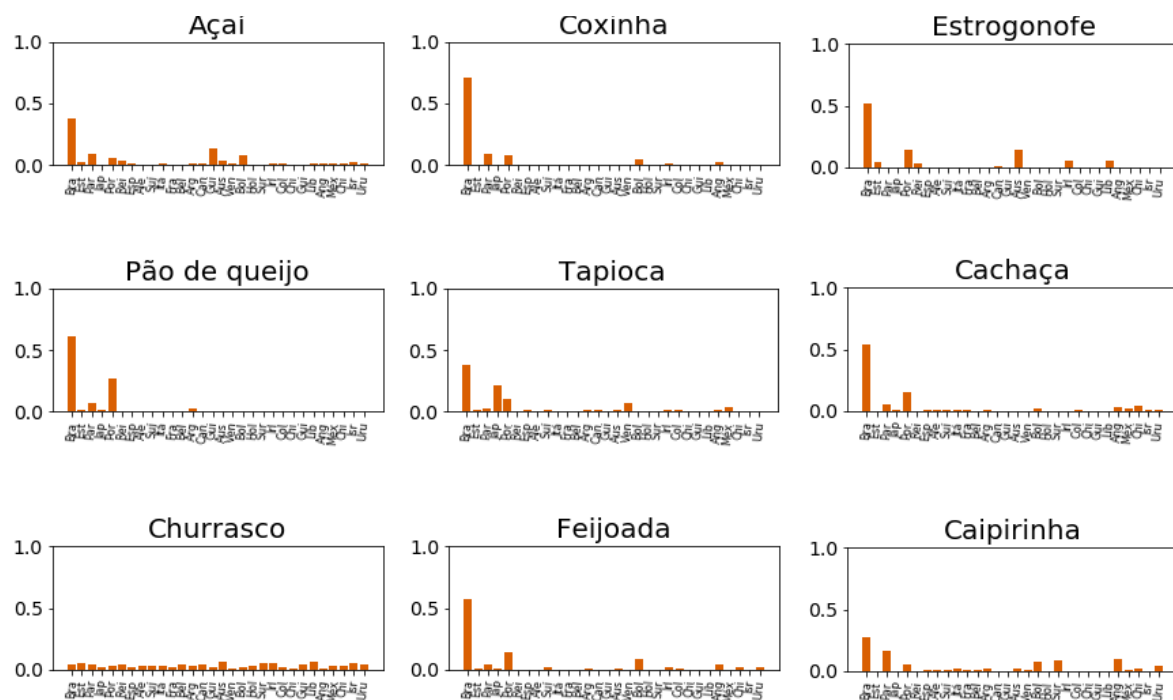


Figure 5.7: Interest focus.

countries, and in fact, it is not only popular in Brazil. Analyzing “Pão de Queijo”, Facebook users from both Brazil and Portugal demonstrate a significant interest in this food. Because of this, the uncertainty increases when we do consider Brazil. This result shows that the interest entropy difference does not depend only on the highest z-score but also considers the whole interest focus distribution.

By using the entropy difference methodology, we create the vectors for each country, by considering now only the interests that are part of the Brazilian cultural identity. Table 5.3 shows correlations between the *Immigrant ranking*, *Expat (Facebook) ranking* and the rankings generated with different measures of distance considering only those 7 interests, and table 5.4 shows the correlation between all of them and the *Geographic weighted distance ranking*. When we compare those rankings considering the Weighted tau correlation, Cosine distance ranking has a 0.40 correlation with the *Immigrant ranking*. Notice that while our methodology allows us to reduce the size of the vectors from 20 to 7 dishes, the correlations between the rankings are kept the same.

As well as the methodology presented to measure the cultural distance, the methodology proposed to identify the interests that are part of the cultural identity of a specific country generates good results when applied to the case study of Brazilian cuisine. The results show that the entropy difference measure can be applied not only

Rankings	WT ⁽ⁱ⁾	KT ⁽ⁱⁱ⁾	S ⁽ⁱⁱⁱ⁾	J ^(iv)
Euclidean distance	0.3786	0.0493 (0.71)	0.0562 (0.77)	0.3333
Cosine distance	0.3897	-0.0985 (0.45)	-0.1374 (0.48)	0.3333
Mean Absolute Error	0.3881	-0.0837 (0.52)	-0.1453 (0.45)	0.4286
Relative Error	0.3861	0.0542 (0.68)	0.0596 (0.76)	0.3333
Earth Mover’s distance	0.3881	-0.0837 (0.52)	-0.1453 (0.45)	0.4286

Table 5.3: Comparison with the *Immigrant ranking*. Considering only the 7 interests typical from Brazil according to entropy difference.

Rankings	WT ⁽ⁱ⁾	KT ⁽ⁱⁱ⁾	S ⁽ⁱⁱⁱ⁾	J ^(iv)
Euclidean distance	0.1998	0.1872 (0.15)	0.266 (0.16)	0.3333
Cosine distance	0.1688	-0.0394 (0.76)	-0.03 (0.88)	0.3333
Mean Absolute Error	0.1755	-0.0049 (0.97)	-0.0074 (0.97)	0.3333
Relative Error	0.1763	-0.0148 (0.91)	0.003 (0.99)	0.4286
Earth Mover’s distance	0.1755	-0.0049 (0.97)	-0.0074 (0.97)	0.3333

Table 5.4: Comparison with the *Geographic weighted distance ranking*. Considering only the 7 interests typical from Brazil according to entropy difference.

to characterize a country but also as an alternative to reduce vector representations. In the next section, we present a general discussion about the main findings in the case study.

5.4 Discussion

In the literature, many measures of distance such as the geographic distance, typically included in gravity-type models of migration, were found useful to characterize similarities between countries. The goal of this study is to explore specific cultural attributes in order to develop a measure of distance that would most accurately characterize cultural affinities between countries with regard to food preferences. By using social media data to characterize the interests of each country enables us to represent it in terms of its cultural composition and to compare the countries by calculating the aforementioned types of distance between them. Such an approach complements previous research that employs various measures of distance in order to explain international migration patterns. The methodology we developed helps one understand the study of cultural attraction from the social media perspective. In addition, the cultural distance between countries can also be included as one additional attribute in classic gravity-type models of migration [Cohen et al., 2008].

The cultural distance can influence migration flows since cultural aspects can

be a factor of attraction between populations. On the other hand, immigration can change the culture of a country. The cultural measure we developed can be included as one additional variable in gravity-type models to explain flows of people around the world. This case study focuses on how Facebook data can be correlated with migration processes and how the Brazilian culture is spread around the world.

Facebook data revealed to be a reliable proxy to study international migration. In this paper, the correlation between the proportion of Facebook users that are Brazilian expats living abroad and the official data about the number of Brazilian immigrants is greater than 0.7. Also, this data can estimate interests related to a foreign culture, in this case, the Brazilian cuisine. The ranking generated considering the similarity between countries given by food interest is almost 0.4 correlated with the ranking generated with the official proportion of Brazilian immigrants in some countries. Our results suggest the cultural similarity between Brazil and some countries occur due to aspects such as geographic proximity (e.g. Paraguay, Argentina), linguistic similarity (e.g. Portugal, Angola), and most importantly, the number of immigrants in the population, which increases the spread of cultural elements from the country of origin. Regarding the Brazilian cultural identity, the results are also optimistic. The entropy difference measure seems to be a good measure not only to characterize a country but also as an alternative to reduce vector representations.

Finally, the next chapter presents more generic conclusions regarding the thesis. We also offer additional comments about the applicability of the results on Gravity-type models for migration, discuss our findings, and future work.

Chapter 6

Conclusion and Future Work

In this chapter, we present a final discussion about the results and offer additional comments about their applicability on Gravity-type models for migration among other applications. In the first section, we also discuss our findings and draw final conclusions and then, in the last section, we mention future work.

6.1 Conclusion

In this thesis, we presented a methodology that leverages the Facebook Advertising Platform to infer the number of people interested in some cultural aspects. Our methodology explores the set of attributes used to define the audience to which an advertiser wants to deliver the advertisement. We focus on identifying the cultural identity of countries by exploring Facebook users' preferences through the data retrieved from the Facebook Advertising Platform and quantifying the cultural distance between countries.

To characterize the cultural identity and measure the cultural distance between countries, we develop a new methodology for exploring the data provided by online social networks on advertising platforms. To decide which interest is part of the cultural identity of each country and therefore characterizes a country and its local population, we made use of an approach that relies on information theory and is based on a measure of entropy. To calculate the cultural distance between countries, the z-score normalization is used to create a vectorized representation of countries in order to compare them with each other. While evaluating the cultural distance between countries, we explore several measures of distance and compare them in the context of cultural affinities.

We then used our methodology to conduct a case study where typical Brazilian dishes are used as a proxy of how the Brazilian culture is consumed across various

countries in the world. In this context, we measure the cultural distance between Brazil and the countries most preferred by Brazilian immigrants. We also apply the methodology to identify dishes that in fact represent the cultural identity of the country.

In this study, Facebook data revealed to be a reliable proxy to study international migration. The correlation between the proportion of Facebook users that are Brazilian expats living abroad and the official data about the number of Brazilian immigrants is greater than 0.7. Also, this data can estimate interests related to a foreign culture, in this case, Brazilian cuisine. The ranking generated considering the similarity between countries given by food interest is almost 0.4 correlated with the ranking generated with the official proportion of Brazilian immigrants in some countries. By selecting only the dishes identified by our methodology as being part of the cultural identity of Brazil, the vectorized representation of countries can be reduced. The results show that our methodology allows us to reduce the vectors while the correlations between the rankings are kept the same.

We should emphasize that, in spite of using the Facebook Advertising Platform as the source of data, our methodology is not limited to the Facebook social media platforms. The methodology may be applied to other online social network platforms, digital advertising systems or even data regarding population preferences from other sources. Our study forms the foundation for many research directions that can be pursued in the future to explore cultural preferences by using representative data, easy to collect with low cost about different populations. Given these characteristics, the data has the potential to be explored in future works.

6.2 Future Work

As future work, we intend to expand our data collection by considering other data sources such as Google Trends, Twitter and LinkedIn. Avoiding being restricted to one single social media will also allow us to compare the results and identify peculiarities across different sources. This new source of data may also help us understand the different public of distinct social networks and study how the differences in the audiences impact their advertising systems.

We also intend to employ our methodology to identify cultural identity from other countries and measure the cultural distance between them. In an optimistic scenario, we expect that our cultural distance measure may be employed as an auxiliary attribute in the classic gravity-type model of migration to explain the flows of people around the world.

The cultural distance can influence migration flows since cultural aspects can be a factor of attraction between countries. On the other hand, immigration can change the culture of a country. Given this relation, the methodology can be replicated to generate results, differentiating the perspective of the Brazilian expat audience relative to the non-Brazilian expat audience. This analysis ought to be interesting as it can provide valuable insights on the extent to which the Brazilian immigrants are responsible for the spread of Brazilian culture around the world.

Bibliography

- [Abbar et al., 2015] Abbar, S., Mejova, Y., and Weber, I. (2015). You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197--3206.
- [Alburez-Gutierrez et al., 2019] Alburez-Gutierrez, D., Zagheni, E., Aref, S., Gil-Clavel, S., Grow, A., and Negraia, D. V. (2019). Demography in the digital era: new data sources for population research.
- [Almerico, 2014] Almerico, G. M. (2014). Food and identity: Food studies, cultural, and personal identity. *Journal of International Business and Cultural Studies*, 8:1.
- [Araujo et al., 2017] Araujo, M., Mejova, Y., Weber, I., and Benevenuto, F. (2017). Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 253--257.
- [Beugelsdijk and Welzel, 2018] Beugelsdijk, S. and Welzel, C. (2018). Dimensions and dynamics of national culture: Synthesizing hofstede with inglehart. *Journal of Cross-Cultural Psychology*, 49(10):1469–1505.
- [Bhugra et al., 1999] Bhugra, D., Bhui, K., Mallett, R., Desai, M., Singh, J., and Leff, J. (1999). Cultural identity and its measurement: a questionnaire for asians. *International Review of Psychiatry*, 11(2-3):244--249.
- [Boutaud et al., 2016] Boutaud, J., Becuț, A., and Marinescu, A. (2016). Food and culture. cultural patterns and practices related to food in everyday life. *International Review of Social Research*, 6:1–3.
- [Brodersen et al., 2012] Brodersen, A., Scellato, S., and Wattenhofer, M. (2012). Youtube around the world: Geographic popularity of videos. In *Proceedings of the 21st International Conference on World Wide Web, WWW '12*, pages 241--250, New York, NY, USA. ACM.

- [Cantarero et al., 2013] Cantarero, L., Espeitx, E., Gil Lacruz, M., and Martín, P. (2013). Human food preferences and cultural identity: The case of aragón (spain). *International Journal of Psychology*, 48(5):881--890.
- [Cesare et al., 2018] Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography*, 55(5):1979--1999.
- [Chunara et al., 2013] Chunara, R., Bouton, L., Ayers, J. W., and Brownstein, J. S. (2013). Assessing the online social environment for surveillance of obesity prevalence. *PloS one*, 8(4).
- [Cohen et al., 2008] Cohen, J. E., Roig, M., Reuman, D. C., and GoGwilt, C. (2008). International migration beyond gravity: A statistical model for use in population projections. *Proceedings of the National Academy of Sciences*, 105(40):15269--15274. ISSN 0027-8424.
- [De Santis et al., 2015] De Santis, G., Maltagliati, M., and Salvini, S. (2015). A measure of the cultural distance between countries. *Social Indicators Research*, 126.
- [Dubois et al., 2018] Dubois, A., Zagheni, E., Garimella, K., and Weber, I. (2018). Studying migrant assimilation through facebook interests. In Staab, S., Koltsova, O., and Ignatov, D. I., editors, *Social Informatics*, pages 51--60, Cham. Springer International Publishing.
- [Faist, 2000] Faist, T. (2000). The volume and dynamics of international migration and transnational social spaces.
- [Fatehkia et al., 2018] Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using facebook ad data to track the global digital gender gap. *World Development*, 107:189 -- 209. ISSN 0305-750X.
- [Fatehkia et al., 2019] Fatehkia, M., O'Brien, D., and Weber, I. (2019). Correlated impulses: Using facebook interests to improve predictions of crime rates in urban areas. *PLOS ONE*, 14(2):1--16.
- [Garcia et al., 2018] Garcia, D., Mitike Kassa, Y., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., and Cuevas, R. (2018). Analyzing gender inequality through large-scale facebook advertising data. *Proceedings of the National Academy of Sciences*, 115(27):6958--6963. ISSN 0027-8424.

- [Ghemawat, 2001] Ghemawat, P. (2001). Distance still matters. *Harvard business review*, 79(8):137--147.
- [Gil-Clavel and Zagheni, 2019] Gil-Clavel, S. and Zagheni, E. (2019). Demographic differentials in facebook usage around the world. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 647--650.
- [Guha et al., 2010] Guha, S., Cheng, B., and Francis, P. (2010). Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 81--87.
- [Head and Mayer, 2002] Head, K. and Mayer, T. (2002). *Illusory border effects: Distance mismeasurement inflates estimates of home bias in trade*, volume 1. Citeseer.
- [Herdağdelen et al., 2016] Herdağdelen, A., State, B., Adamic, L., and Mason, W. (2016). The social ties of immigrant communities in the united states. In *Proceedings of the 8th ACM Conference on Web Science, WebSci '16*, pages 78--84, New York, NY, USA. ACM.
- [Isard, 1954] Isard, W. (1954). Location theory and trade theory: short-run analysis. *The Quarterly Journal of Economics*, pages 305--320.
- [Knight, 1966] Knight, W. R. (1966). A computer method for calculating kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 61(314):436-439.
- [Kogut and Singh, 1988] Kogut, B. and Singh, H. (1988). The effect of national culture on the choice of entry mode. *Journal of International Business Studies*, 19(3):411--432. ISSN 1478-6990.
- [Kosinski et al., 2015] Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543.
- [Laakso and Taagepera, 1979] Laakso, M. and Taagepera, R. (1979). "effective" number of parties: a measure with application to west europe. *Comparative political studies*, 12(1):3--27.
- [Massey et al., 1993] Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., and Taylor, J. E. (1993). Theories of international migration: A review and

- appraisal. *Population and Development Review*, 19(3):431--466. ISSN 00987921, 17284457.
- [Mayer and Zignago, 2011] Mayer, T. and Zignago, S. (2011). Notes on cepii's distances measures: The geodist database.
- [Mejova et al., 2018a] Mejova, Y., Gandhi, H. R., Rafaliya, T. J., Sitapara, M. R., Kashyap, R., and Weber, I. (2018a). Measuring subnational digital gender inequality in india through gender gaps in facebook use. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1--5.
- [Mejova et al., 2018b] Mejova, Y., Weber, I., and Fernandez-Luque, L. (2018b). Online health monitoring using facebook advertisement audience estimates in the united states: evaluation study. *JMIR public health and surveillance*, 4(1):e30.
- [of Migration, 2017] of Migration, I. O. (2017). World migration report 2018.
- [Palotti et al., 2020] Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Garcia Herranz, M., Al-Asad, M., and Weber, I. (2020). Monitoring of the venezuelan exodus through facebook's advertising platform. *PLOS ONE*, 15(2):1--15.
- [Pöttschke and Braun, 2017] Pöttschke, S. and Braun, M. (2017). Migrant sampling using facebook advertisements: A case study of polish migrants in four european countries. *Social Science Computer Review*, 35(5):633--653.
- [Rama et al., 2020] Rama, D., Mejova, Y., Tizzoni, M., Kalimeri, K., and Weber, I. (2020). Facebook ads as a demographic tool to measure the urban-rural divide. *arXiv preprint arXiv:2002.11645*.
- [Rampazzo et al., 2018] Rampazzo, F., Zagheni, E., Weber, I., Testa, M. R., and Bil-lari, F. (2018). Mater certa est, pater numquam: What can facebook advertising data tell us about male fertility rates? In *Twelfth International AAAI Conference on Web and Social Media*.
- [Recchi and Favell, 2019] Recchi, E. and Favell, A. (2019). *Everyday Europe: Social transnationalism in an unsettled continent*. Policy press.
- [Ribeiro et al., 2018] Ribeiro, F. N., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth AAAI Conference on Web and Social Media, ICWSM'18, Stanford, USA*.

- [Ribeiro et al., 2019] Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Oana Goga, F. B., Gummadi, K. P., and Redmiles, E. M. (2019). On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *ACM Conference on Fairness, Accountability, and Transparency, FAT*’19*, Atlanta, USA.
- [Sajadmanesh et al., 2017] Sajadmanesh, S., Jafarzadeh, S., Ossia, S. A., Rabiee, H. R., Haddadi, H., Mejova, Y., Musolesi, M., Cristofaro, E. D., and Stringhini, G. (2017). Kissing cuisines: Exploring worldwide culinary habits on the web. In *Proceedings of the 26th international conference on world wide web companion*, pages 1013–1021.
- [Sibal, 2018] Sibal, V. (2018). Food: Identity of culture and religion. 6:10908–10915.
- [Silva et al., 2020] Silva, M., de Oliveira, L. S., Andreou, A., de Melo, P. O. V., Goga, O., and Benevenuto, F. (2020). Facebook ads monitor: An independent auditing system for political ads on facebook. *arXiv preprint arXiv:2001.10581*.
- [Silva et al., 2014] Silva, T. H., de Melo, P. O. V., Almeida, J. M., Musolesi, M., and Loureiro, A. A. (2014). You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare. In *Eighth International AAAI Conference on Weblogs and Social Media*.
- [Simpson, 1949] Simpson, E. H. (1949). Measurement of diversity. *nature*, 163(4148):688–688.
- [Smith et al., 2018] Smith, A., Anderson, M., et al. (2018). Social media use in 2018. *Pew research center*, 1:1–4.
- [Spyratos et al., 2018] Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2018). Migration data using social media: a european perspective. *Publications Office of the European Union*.
- [Spyratos et al., 2019] Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2019). Quantifying international human mobility patterns using facebook network data. *PloS one*, 14(10).
- [Stewart et al., 2019] Stewart, I., Flores, R. D., Riffe, T., Weber, I., and Zagheni, E. (2019). Rock, rap, or reggaeton?: Assessing mexican immigrants’ cultural assimilation using facebook data,. In *The World Wide Web Conference, WWW ’19*, pages 3258–3264, New York, NY, USA. ACM.

- [Tung and Verbeke, 2010] Tung, R. L. and Verbeke, A. (2010). Beyond hofstede and globe: Improving the quality of cross-cultural research.
- [Vieira et al., 2020] Vieira, C. C., Ribeiro, F. N., de Melo, P. O. S. V., Benevenuto, F., and Zagheni, E. (2020). Using facebook data to measure cultural distance between countries: the case of brazilian cuisine. In *The Web Conference, WWW '20*.
- [Vigna, 2015] Vigna, S. (2015). A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web*, pages 1166--1176.
- [Wagner et al., 2014] Wagner, C., Singer, P., and Strohmaier, M. (2014). Spatial and temporal patterns of online food preferences. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 553--554.
- [Weber et al., 2018] Weber, I., Kashyap, R., and Zagheni, E. (2018). Using advertising audience estimates to improve global development statistics. *Itu Journal: Ict Discoveries*, 1(2).
- [West et al., 2013] West, R., White, R. W., and Horvitz, E. (2013). From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1399--1410.
- [Zagheni et al., 2017] Zagheni, E., Weber, I., Gummadi, K., et al. (2017). Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43(4):721--734.
- [Zwillinger and Kokoska, 1999] Zwillinger, D. and Kokoska, S. (1999). *CRC standard probability and statistics tables and formulae*. Crc Press.