Mehrjoo et al. *EPJ Data Science* (2022) 11:40
https://doi.org/10.1140/epjds/s13688-022-00350-w

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

A new methodology to measure faultlines at scale leveraging digital traces

Amir Mehrjoo[1,2*], Rubén Cuevas[2,3] and Ángel Cuevas[2,3]

*Correspondence:
amir.mehrjoo@imdea.org
[1]IMDEA Networks Institute, Leganes, Spain
[2]Telematics Engineering Department, Universidad Carlos III of Madrid (UC3M), Madrid, Spain
Full list of author information is available at the end of the article

Abstract

The definition of society is tight with human group-level behavior. Group *faultlines* defined as *hypothetical lines splitting groups into homogeneous subgroups based on members' attributes* have been proposed as a theoretical method to identify conflicts within groups. For instance, crusades and women's rights protests are the consequences of strong faultlines in societies with diverse cultures.

Measuring the presence and strength of faultlines represents an important challenge. Existing literature resorts in questionnaires as traditional tool to find group-level behavioral attributes and thus identify faultlines. However, questionnaire data usually come with limitations and biases, especially for large-scale human group-level research. On top of that, questionnaires limit faultline research due to the possibility of dishonest answers, unconscientious responses, and differences in understanding and interpretation.

In this paper, we propose a new methodology for measuring faultlines in large-scale groups, which leverages data readily available from online social networks' marketing platforms. Our methodology overcomes the limitations of traditional methods to measure group-level attributes and group faultlines at scale.

To prove the applicability of our methodology, we analyzed the faultlines between people living in Spain, grouped by geographical regions. We collected data on 67,270 interest topics from Facebook users living in Spain, France, Germany, Greece, Italy, Portugal, and the United Kingdom. We computed existing metrics to measure faultlines' distance and strenght using our data to identify potential faultlines existing among Spanish regions. The results reveal that the strongest faultlines in Spain belong to Spanish Islands (the Canary Islands and the Balearic Islands), Catalonia, and Basque regions. These findings are aligned with the historical secessionist movements and cultural diversity reports supporting the validity of our methodology.

Keywords: Conflict measurement; Social network analysis; Culture; Faultlines; Subgroups; Advertising data

1 Introduction

"Conflict is the beginning of consciousness" – M. Esther Harding. A short period after settling the flames of World War II, many social scientists started thinking about how to explain the psychological forces that culminated in the Holocaust, among other horrors. During the post-war period leading into the 1970s, a branch of social scientists focused

Springer

on group and group-formation procedures to find an interpretation for conflicts related to collective human behavior. In this context, 'group' was a label for aggregated interpersonal processes. Measurement techniques for the group-level behavior lack consistent findings when considering single group members' attribute such as race [1–6]. Therefore, researchers were motivated to investigate the impact of multiple group member attributes alignment (e.g. race and gender) on team members conflicts.

Faultlines are hypothetical dividing lines splitting a team into one or more relatively homogeneous subgroups [7, 8]. Studies on the effects of faultline dynamics to explain theoretical underpinnings and effects of faultlines appear in sociology literature [9, 10]. Like many other aspects of human behavior, the implementation of measurement tools has been challenging. Still, reliable measurement techniques associated with group-level attributes have been introduced by the literature [11, 12].

Group faultlines usually have a detrimental effect on team-level outcomes [8, 10]. Lau and Murnighan (1998) introduced the initial faultline theoretical model [7, 13]. They based the theoretical reasoning on social categorization and social identity approaches [14]. Despite a well-developed theoretical framework, limited measurement techniques currently exist to create a strong link between these theories and the real world. Managers and politicians have considered faultlines measurements an essential tool for managing performance and leadership. Thanks to technological developments during the past few decades, many aspects of human social behavior are now more apparent to scientists. One of the main contributions of technology to human life is the onset and spread of social network platforms. These platforms offer free services to users in exchange of access to users' data; they enrich their databases by the behavioral attributes of their users and manipulate them for marketing purposes. For example, Facebook provides a marketing platform for advertisers to target their audiences based on demographic, behavioral characteristics and location. These platforms' new social behavior measurement instruments have more valuable benefits than traditional ones (surveys and questionnaires).

The traditional approach to measure faultlines was the application of questionnaires by asking team members about their behavioral attributes and calculating the metrics. This approach exposes the results to biases such as dishonest or unconscientious responses. Besides, scaling the research to larger groups using this approach is costly and time-consuming. In this research, we employ data from social networks' marketing platforms and introduce a new approach to overcome these limitations. This new approach aims to increase the scalability and accuracy of faultlines measurement while making it less expensive. We introduce a reliable methodology based on data from billions of social network users to measure the faultline separating populations in different geographic regions. To prove the applicability of this tool, we analyzed the faultlines between people living in Spain, grouped by geographical regions. Spain has experienced identity-related regionalism independence movements and conflicts. If our methodology performs well, it should be able to capture these conflicts.

2 Theoretical discussion

The regional/national identity salience in geographic regions produces the previously mentioned conflicts. Political leaders tend to promote the differences such as cultures and national identities for getting more votes by drawing a clear line (faultline) between them and us (e.g., Catalans vs. non-Catalans in Spain). Social science literature is rich in theories

and measurement techniques to analyze faultlines. We extend the available measurement techniques to understand better the faultlines' status of large-scale groups. This research proves how our new proposed technique measures well faultlines between groups living in different Spanish Autonomous Communities[1] (referred to as CCAAs by their abbreviation in Spanish).

We first apply self-categorization and social identity theories to identify the places where we expect to find strong faultlines. Then, we use one of the most popular online social network platforms (Facebook) to measure the faultline's distance and strength.

2.1 Faultline theories

The term faultline originates from geography and refers to the intersection of two tectonic plates. Therefore, faultlines mark locations that are more prone to split. Lau and Murnighan (1998) adopted this definition for research in group conflicts by defining faultlines as "hypothetical dividing lines that may split a group into subgroups based on one or more attributes" [7]. The purpose behind measuring faultlines is to quantify how a team is prone to split into subgroups [15]. According to the faultline theories, the groups divided into two homogeneous subgroups with distant intra-group attributes are more likely to conflict between members. Three main categories of faultlines have been the focus of the articles in this literature (1) Separation-based faultlines (e.g., followers of different football teams) (2) Information-based faultlines (e.g., engineers vs. psychologists) (3) Resource-based faultlines (group members' access to "finite resources, e.g., power, materials, authority, and status") [16]. Social identity and self-categorization are two of the most prevailing theories in this field. They are building blocks for faultline research as they explain: (1) sub-group formation, (2) relationship between group identity and trust, and (3) the nature of ingroup-outgroup biases [17].

2.2 Self-categorization and social identity

Social categorization theory justifies faultlines in human groups, and the comparative-fit is one of the several factors affecting social categorization processes. Comparative-fit explains how observed similarities and differences, such as languages or accents, are perceived as social categories [14]. A strong faultline makes the differences within groups more salient. The human's brain ability to process information is limited. For example, if we see an object is flying and singing, we unconsciously assume it is a bird. Then we assign all the bird category attributes such as breeding by laying eggs and having wings to that object. Therefore, abstraction is the key attribute of the human brain to understand the surrounding world. The legitimate model of the human brain is the highest level of abstraction for demonstrating cognitive mechanisms [18].

The human capacity to recognize different levels of abstraction is limited. Cognitive procedures such as abstraction, thinking, and learning structure the information we retrieve from the world outside. When individuals confront disorganized and unlabelled data, they abstract the complex data into basic concepts with specific goals [18]. If the flying object has one wing instead of two in the bird example, the human's mind still puts it in the birds' category. The same happens when we see someone speaking a language (Italian,

[1] Autonomous Communities refer to the first level administrative division in Spain. They are equivalent to states in United States or landers in Germany.

Chinese, Catalan, etc.), then we unconsciously assign attributes to that person (e.g., the origin country is Italy, China, Catalonia, etc.). The social identity approach describes the state of people thinking of themselves and others as a group. This theory states the three steps of psychological processes to perceive the social group is: (1) social categorization: organize social information by categorizing people into groups such as Catalan, Castilian, South American, and Japanese. (2) social comparison: give meaning to those categories to understand the group's task in the specific situation (e.g., Catalans speak Catalan, Japanese are hardworking). (3) social identification: the process in which people relate themselves to one of those categories (e.g., I am Catalan!, I am Spanish!).

The lowest level of abstraction is given as a personal self during this process, where the perceiver categorizes themself as "I". A higher level of abstraction corresponds to a social self, where the perceiver categorizes themself as part of a "we" compared to a salient out-group (them) [19]. Social identity theory explains some behavioral attributes of group members. According to this theory, people maintain their self-esteem by a cognitive bias assigning positive attributes to their group, nationality, category, etc. Individuals are assumed to be intrinsically motivated to achieve positive distinctiveness and "strive for a positive self-concept" [20]. This cognitive bias may also result in uneven distribution of resources and discrimination within groups. Therefore, members endorse resource distributions that would maximize the positive distinctiveness of an in-group in contrast to an out-group at the expense of personal self-interest [14]. Self-identity theory also explains that an in-group seeks to increase self-esteem by direct competition against the out-group. This effect would cause polarization of the group at a high level of social competition and make two salient subgroups. According to the similarity attraction paradigm [21], members in one subgroup experience psychological distance from other subgroup members and are less likely to cooperate [22]. Therefore, people living in the same country feel they are in the same group, thus, they have less distant behavioral attributes than the people in other countries (other groups) (*Proposition 1*). The self-categorization theory argues that a category's prototype is contingent on the context in which the category is encountered. This theory is consistent with leader categorization theory, whereby stereotypical leaders were more effective than non-stereotypical leaders [23].

2.3 Insular effects

Islands have developed isolated living communities, whether plant, animal, or human, separated from, and differing to varying degrees from, mainland communities of the same kind. Means of physical communications, such as transport, were crucial for the past interaction of island and continental populations. They were also largely dependent on distance from the mainland, the climate, and technology. Contacts are influential in determining the degree, and the nature of cultural factors [24]. This is especially true in islands, which have been less affected by the cultural and ethnic change, hostile invasion, mass immigration, or political interference, and at the same time have been more exposed, if not open, to cultural stimuli from a wider variety of sources [25]. The distance and insularity of these islands result in more differential cultural attributes in the population. The differential cultural attributes may grew a strong regional identity and made it prevailing compared to the countries' national identity. According to the faultline theories, the inhabitants of islands should consider themselves a distinct group that will lead to a strong faultline. Therefore, the faultlines in the islands are expected to be relatively strong (*Proposition 2*).

2.4 Conflict

Consensus exists in the faultline theories literature that a strong, activated faultline in a group of people can explain some social conflicts (task conflict, relationship conflict, process conflict) [26]. Faultline activation is defined as the process by which members of a group are perceived as members of one or more subgroups [27]. A vast body of literature is devoted to developing theories and techniques for measuring and managing conflicts (e.g., international joint ventures [28, 29], bi-cultural family kids [30]). The existing literature considered a strong faultline an important predictor of group conflicts. Thatcher and Patel (2011) argued that if a group perceives other sub-groups as threatening, individuals maintain their self-esteem by positive distinctiveness, resulting in a conflict between subgroups [8]. On the other hand, group diversity decreases conflict and group faultline strength [31].

Therefore, based on the literature outcomes, our methodology should observe higher faultlines values in regions that experienced some regional conflicts in their history (*Proposition 3*).

2.5 Measurement

The literature in the human group-level measurements mainly relies on questionnaire surveys. The application of questionnaires in the analysis comes with limitations, such as the number of questions and non-scalability. Large-scale surveys and collecting empirical data on the population have been costly, time-consuming, and in many instances impossible during human history [32]. On the other side, advertisers can elicit many behavioral dimensions by tracking internet users' online behaviors [33]. Such platforms continuously track users' interests, beliefs, preferences, behaviors, locations, and interactions. The majority of faultline research has been conducted by questionnaire survey-based experiments using relatively small groups. This paper is the first attempt at using large-scale field data provided by online social platforms in faultline research. We use one of the most prominent social networking platforms' data (Facebook), with more than 2.9B monthly active users, to measure the faultlines. Facebook places particular importance on classifying the interests of its users for marketing purposes [34] and measures all the individual user's preferences.

2.5.1 Interests in Facebook

Facebook infers user preferences from self-reported interests, clicking behaviors on Facebook posts, software downloads, GPS location, and processing the communications with other users in multiple platforms (e.g., Facebook, Instagram or Whatsapp). Facebook makes this information anonymized and accessible to marketers through an application programming interface (API). Facebook finds users' interests by tracking their activities on Facebook's platforms (i.e., Facebook and Instagram) and third-party websites, apps, and online services. To be more specific, in addition to the information collected from its owned social networks and applications (Facebook, Instagram, and Whatsapp), Facebook collects data from more than 30% of the most popular websites [35].

Facebook may also track users' locations through their mobile devices, inferring the amount of time each user spends in locations such as football fields, universities, theaters, restaurants, and churches. Facebook users' interests are shaped by multiple facets of their activity (e.g., if someone goes to the football stadium for all Real Madrid football

team matches, after checking out Real Madrid online website, Facebook most probably assigns "Real Madrid football team" to the user interests). Thus, countless interests shape human preferences in Facebook. Facebook organizes interests in a multi-level, hierarchical structure with 14 root categories: business and industry, education, family and relationships, fitness and wellness, food and drink, hobbies and activities, lifestyle and culture, news and entertainment, people, shopping and fashion, sports and outdoors, technology, travel places, and events. Facebook also assigns unique, language-independent id to each interest.

Facebook finds user interests through multiple information channels, including page likes, self-declared interests, downloaded apps, and location. This approach forms the most comprehensive dictionary of preferences for billions of people. Previous studies found the following paths to assign preferences to each user. The user has this preference because: (i) "This is a preference the user added," (ii) "what the user does on Facebook, such as pages the user has liked or ads the user clicked," (iii) "the user clicked on an ad related to…", (iv) "the user installed the app…", (v) "the user liked a page related to…", (vi) "the user comments, posts, shares or reactions the user made related to…". [36] The goal here is to measure faultlines (strength and distance) using the features extracted from different the popularity of different topics among groups of people living in specific geographic regions. The following section explains how we extracted these features from Facebook data. In previous work, we presented the first large-scale analysis of measuring culture using tens of thousands of interests to define human group culture and examined the validity of this approach using the world values survey (WVS), among other sources. Our findings showed that the Facebook measurement encompasses a broader range of cultural explanatory dimensions than the WVS [37].

2.5.2 Faultline distance

According to distance theory, team members in one subgroup feel psychological distance from team members in other subgroups, making them less likely to cooperate [22]. Thus, measuring the distance between the behavioral attributes of the subgroups will shed light on the status of the faultline. Faultline distance reflects the extent to which formed subgroups differ from one another in terms of behavioral characteristics [38]. The distance between the group-level attributes of two subgroups is used to calculate faultline distance. Consider group G consists of n members A_j ($j = 1$ to n).

$$G = \{A_1, A_2, \ldots, A_n\}.$$

Each member of the group may be interested in topic i ($a^i = 1$) or may not be interested in that topic ($a^i = 0$). Then we can assign a vector of p dimensions (attributes) to each member (e.g. member j).

$$\overrightarrow{A}_j = \left\langle a_j^1, a_j^2, \ldots, a_j^p \right\rangle.$$

We compute group-level attributes ($\overrightarrow{V}_g$) using mean vectors (average value of group members for each attribute). The pth group level attribute ($\overline{a^p}$) is calculated by averaging

pth attribute ($\overline{a_j^p}$) across all group members (n).

$$\overline{a^p} = \frac{\sum_{j=1}^{n} a_j^p}{n},$$
$$\overrightarrow{V}_g = \langle \overline{a^1}, \overline{a^2}, \ldots, \overline{a^p} \rangle.$$

Faultlines by definition are hypothetical lines splitting group (V) into subroups (v_1, v_2). We assign a vector of p dimensions to each subgroup ($\overrightarrow{v_1}, \overrightarrow{v_2}$):

$$\overrightarrow{v_1} = \langle \overline{a_1^1}, \overline{a_1^2}, \ldots, \overline{a_1^p} \rangle,$$
$$\overrightarrow{v_2} = \langle \overline{a_2^1}, \overline{a_2^2}, \ldots, \overline{a_2^p} \rangle.$$

The faultline distance (D_g) is the Eclucidian distance between two subgroup attribute vectors (v_1, v_2):

$$D_g = |\overrightarrow{v_1} - \overrightarrow{v_2}| = \sqrt{\sum_{i=1}^{p} \left(\overline{a_1^i} - \overline{a_2^i}\right)^2}.$$

2.5.3 Faultline strength

Thatcher and Patel (2003) described faultlines as potential splits that yield "relatively homogeneous subgroups based on the attributes of the team members." [39]. Faultlines, as the definition implies, are imaginary lines that separate homogeneous groups, and faultline strength measures how homogeneous these subgroups are. As a result, to calculate faultline strength, referred to as *Fau*, we compute the variations within each group. This measurement is based on self-categorization theory, which distinguishes between in-group and out-group, which explains why the measure can detect only two subgroups.

In theory, polarization is one outcome of group conflict, making within-group differences more salient [40]. Therefore, faultline strength is a valid measurement for groups with strong faultlines. They illustrated the differences between faultline and distance measurement using a comparison table. The Table 1 shows two groups of four people with different demographics. In the first group, there are two distinct subgroups with demographic characteristics that are homogeneous within subgroups. Members of the second group, on the other hand, have a wide range of demographic characteristics. These two groups have the same faultline distance measurement. However, due to the demographic attributes alignment of the subgroup members, the faultline strength measurement of the first group is higher. Thatcher formulated the faultline strength based on information on

Table 1 Example of Subgroup Distance/Fau Strength Analysis (Adopted from Thatcher 2003 [39])

Subgroup 1		Subgroup 2		Dist.	Fau
Memb. 1	Memb. 2	Memb. 3	Memb. 4		
White French	White French	Black British	Black British	0.7	1
White French	Brown Italian	Black British	Asian Chinese	0.7	0.3

p attributes of each group member as follows:

$$Fau_g = \frac{\sum_{i=1}^{p} \sum_{j=1}^{2} n_j^g (\overline{a_j^i} - \overline{a^i})^2}{\sum_{i=1}^{p} \sum_{j=1}^{2} \sum_{k=1}^{n_j^g} (a_{jk}^i - \overline{a^i})^2}, \quad g = 1, 2, \ldots.$$

In this equation, p denotes the number of attributes in the data, n_j^g is the count of members in the subgroup j divided by the split g (we assume the faultlines each time splits the group into two subgroups), $\overline{a_j^i}$ is the mean value of attribute i in subgroup j, $\overline{a^i}$ is the mean value of attribute i in the whole group and a_{jk}^i is the ith attribute of the kth member in subgroup j. *Fau* can take values between 0 and 1, corresponding to faultline strength [39]. Groups split into two relatively homogenous subgroups will have larger *Fau* values. Bezrukova et al. (2009) introduced a new faultline measurement by multiplying faultline distance and faultline strength (Fau), which is more explanatory than the previous measurements [38].

$$S_{\mathrm{Be}} = Fau_g \times D_g.$$

3 Case study (Spain)

With 505,990 square km, Spain is the second-largest country in the European Union, including regions in Africa and several islands in the Atlantic Ocean and the Mediterranean Sea. Spain is divided into 17 autonomous communities (CCAA) regions with six official languages. Spain's continental European territory is located on the Iberian Peninsula. Its insular territory includes the Balearic Islands in the Mediterranean sea and the Canary Islands in the Atlantic ocean in front of Morocco. The minimum Euclidean distance between continental Spain and the Balearic and Canary islands is 87 and 1701 kilometers, respectively. People in Spain live under the same flag and share the same resources. Therefore, they consider themselves belonging to the same group and people in other countries belonging to a different group. According to the theoretical discussion we had in Sect. 2.2, which led to Proposition 1, people living in Spain should have less distant behavioral attributes among them than with the people in other countries (H1).

3.1 Traces of recent conflicts in Spain

Cultural differences and increased interaction among Spanish people as a result of transportation improvements result in regional conflicts and, as a result, radical independence movements in recent history.

3.1.1 Basque region conflict

The leaders of this region's independence movements began political activities in 1958 with the formation of the Euskadi 'ta Askatasuna (ETA) group. The activities of the ETA were organized around four pillars: political, cultural, military, and economic. Throughout its history, the ETA has had numerous conflicts with the national government. The previous regime (General Franco's dictatorship) imprisoned many ETA members as the strategy of action-repression-action took hold. Assassinations against Franco's prime minister, Luis Carrero Blanco, began in December 1973. Two years later, Koordinadora Abertzale Sozialista (KAS) established itself as a coordinating body of this group by executing two members of the political-military branch, prompting widespread condemnation within

the Basque region. Later that year, in 1976, KAS introduced a platform of minimum conditions for MLNV (Basque national liberation movement) participation. This umbrella term encompassed all social, political, and armed organizations based on ETA ideas. ETA's illegal military activities claimed the lives of 92 people in 1980. ETA terrorist attacks continued until 2011, but radical movements continued until 2014. [41].

3.1.2 Catalonia region conflict

In 2003 and 2004, the political situation in the Catalonia region changed quite dramatically. For the first time in the constitutional period, the national and Catalan governments coincided with the same party (the Socialist party). In 2013, openly secessionist parties successfully owned the majority of the seats in the regional parliament (74 seats out of 135). Secessionist parties successfully got more than 49 percent of the popular vote supported. Moreover, according to most polls in 2013, 55 percent of Catalans wanted an independence referendum, and 43% would vote for complete secession from Spain [42].

In October 2017, the Catalonia regional government unilaterally declared independence from Spain. This event triggered: (i) a political reaction from the elected Spanish government that dismissed the Catalonian government, overtaken its responsibilities, and announced new regional elections by December 2017. (ii) a legal reaction that led to the imprisonment of a significant part of the Catalonian government under the accusation (among others) of breaking the Spanish Constitution (iii) the exile of the Catalonian prime minister as of October 2017 and part of its government to different European countries. This dispute resulted in an activated strong faultline dividing the Spanish society into two parts: (i) the ones supporting the right of Catalan people to become an independent country (ii) the ones supporting the national government arguing that all the Spanish citizens have the right to decide about the destiny of Catalonia.

3.1.3 Insular effects in Spain

Spain is a vast country with islands distant from the other CCAAs located in the Iberian Peninsula. The Canary Islands are the southernmost of the Spanish CCAAs and are located in the African Tectonic Plate. The closest part of these Islands to Africa is only 100 kilometers away from west of Morocco. The results from a survey on territorial self-identification conducted between 1990 and 1995 show that the regional identities are strongest due to cultural specificity in the Basque Autonomous Region (49.7%), the Canary Islands (47.9%), and Catalonia (31.8%). The other CCAAs like Galicia (24.2%) and Andalucía (20%) show less prevalent regional identity. The surveys also reveal that nearly half the population feels only Basque, Canarian, and Catalan than Spanish. The insularity (the Islands) and distinctive lifestyle are some of the cultural specificities which cause this attitude among Spaniards [43, 44]. As a result these Islands ran through deadly independence movements e.g. MPAIAC.[2] Therefore, we expect the faultline strength in this region has a relatively high value.

Based on the theoretical discussion we had in Sect. 2.3, we should find the strongest faultlines in Spanish Islands (Canary and Balearic) (Proposition 2) and the regions which experienced secessionist conflicts (Proposition 3) (Catalonia and Basque country) in their history (H2).

[2] Movimiento por la Autodeterminación e Independencia del Archipiélago Canario.

4 Methodology

4.1 Spain as a supergroup

We consider Spain a supergroup for all populations grouped by CCAAs, with a unique national identity, sports teams, currency, and history. Therefore, we expect more alignment between interests and behavioral attributes between Spanish people than between Spanish/non-Spanish people. By implication of the theory and measurement technique, we introduced in the previous section, the distance between behavioral attributes of people inside Spain should be less than the distance between Spanish/non-Spanish people. Polzer et al. (2006), by a study on 45 teams consisting of members from 10 different countries, theorized that geographically dispersed teams are likely to activate faultlines [29]. They also found that these faultlines were stronger when dividing the group into two equally sized subgroups with homogeneous nationality. In Spain, the regional identities of the population in different geographically separated CCAAs are strong. Populations in geographically distant regions such as islands develop stronger regional identities and faultlines due to the physical separation and insularity. Besides, the history of independent movements and political leadership activism in the Catalonia and Basque regions enhance the self-categorization process, leverage regional identity, and strengthen faultlines between regional/non-regional people. Therefore, we hypothesize that the strongest faultlines in Spain separate the population in the Islands (due to geographical distance) and Catalonia, and Basque regions (due to conflict history).

4.2 Application of marketing data

To measure the faultline distance using the Facebook marketing platform, we constructed a vector consisting of the popularity of interests in each region. We calculated the interest penetration (*IP*) for interest i in geographic region b using the following formula:

$$IP_b^i = \frac{MAU_b^i}{MAU_b^{\text{total}}}.$$

In the previous equation, MAU_b^i is the monthly active Facebook users who are interested in the topic i in region b and MAU_b^{total} is the total monthly active Facebook users in region b.

We constructed Interest penetration vector (IPV_b^m) using m interest for the geographic region b:

$$IPV_b^m = \left(IP_b^{\text{topic}^1}, IP_b^{\text{topic}^2}, \ldots, IP_b^{\text{topic}^m}\right).$$

We consider IPV_m^b a proxy for the group level attributes of people living in geographic regions b. We calculated the cosine distance between behavioral attributes of people living in geographic regions k and l using the following equation: The cosine distance will be a value between 0 and 1, where 0 yields the least cosine distance between two regions.

To compare the regions at the first place we used cosine distance of the IPVs which is calculated using the following equation:

$$Distance = 1 - \frac{IPV_k \cdot IPV_l}{\|IPV_k\| \|IPV_l\|}.$$

In the previous equation, k and l are referring to region k and region l respectively. We calculated the Bezrukova faultline strength measurement using the Euclidean distance between the two vectors [38]. We use the following formula to calculate the Euclidean distance between two *IPVs*:

$$D_g = |\overrightarrow{IPV_k} - \overrightarrow{IPV_l}|,$$

$$D_g = \sqrt{\sum_{i=1}^{p} \left(IPV_k^{\text{topic}^i} - IPV_l^{\text{topic}^i}\right)^2}.$$

5 Dataset

We aim at creating a dataset that allows us to construct vectors $\overrightarrow{V}$ for a large number of dimensions for each of the Spanish regions (and other countries) considered in our work. To this end, we will rely on the Facebook's marketing API[3] that allows retrieving the number of users interested in a particular element in a given geographical area (e.g., country, region, etc.).

Marketers use the configurations available in the Facebook platform to target the relevant audiences for their campaigns. Facebook ads manager is a public interface for any Facebook user that allows marketers to define the group they want to target. The group specifications can include the geographic location (e.g., country, region, city, and zip code), demographics (e.g., gender, age, language, and education), behaviors (e.g., mobile device, operating system, and browser), and interests (e.g., sports, food, cars, and art).

The Facebook marketing API provides monthly (MAU) and daily (DAU) active users and relevant advertising costs for a given set of audience specifications. For our research, we listed a set of interests for each set of (interest × region) and sent programmatic queries to Facebook to retrieve monthly and average daily users. Facebook provides three types of location to target specific people (recent location, home location, and travel in). From these options, we chose the users' home location because it is a more reliable and permanent way to identify and locate users who use both mobile and desktop computers. To correctly identify a user's home location, Facebook employs a number of techniques, including information based on IP address, current city in user's profile, and friends' declared profile locations. The data collecting task was performed between February 10th and February 12th, 2021. We evaluated another sample taken between February 25th and February 27th, 2021, to confirm the validity of the results given in the paper is stable in both considered data samples.

The goal to accomplish our research is being able to define a very long vector including tens of thousands of elements with all sorts of information that captures the culture and interests of a region (from food or sports preferences to religion or political issues).

Unfortunately, Facebook does not provide a comprehensive list of all the available interests. Therefore, we had to define our own agnostic mechanisms to create the refereed interest vector. We defined two methods to create vectors, including tens of thousands of interests. Comparing the results obtained with these two methods, we can confirm that the results of our work remain the same as long as the interest vector includes a large enough sample of interests. Following, we describe both interests' definition methods:

[3] developers.facebook.com/docs/marketing-api

- *DBPedia*: We rely on the interest vector used in our previous work [37]. We downloaded 12,301,672 entries from DBPedia (http://es.dbpedia.org/), mapped them into 399,182 Facebook interests, and selected only those linked to audiences with more than 500K FB users worldwide. This process led to a vector of 77,523 interests.
- *FDVT*: To obtain the second vector, we retrieved all the interests assigned to Spanish users who downloaded our online app *Facebook data valuation tool (FDVT)*[45]. FDVT is a tool that informs FB users of an estimation of the revenue they generate for FB based on the targeted ads they visualize and click on. Using the FDVT, we collect FB's interests from users who have installed the tool. Overall, we collected 67,270 interests from 2101 FDVT Spanish users. While we acknowledge those users are not a representative sample of Spanish users, they are enough to achieve our goal of creating a very long vector. In addition, in this method, many interests are likely related to specific regional elements.

The cosine distance between any pair of regions considered in the paper deference was less than 0.001 when considering DPpedia vs. FDVT methods. As a result, because the FDVT interest vector was obtained directly from Spanish users, we conduct our analysis and measurement with such dataset. We create one interest vector per CCAA using the FDVT interest vector (Spain has 17 CCAAs). This collection process yielded a vector of 67,270 interests for each CCAA, including the monthly active users (MAU) interested in each of these 67,270 interests. We also counted the total number of MAU in each CCAA. The MAU of each interest was then divided by the MAU of all Facebook users in each CCAA, yielding a value ranging from 0 to 1, which we used as a proxy for the behavioral characteristics of people living in each CCAA. Note that we obtain the same data for six European countries (France, Germany, Greece, Italy, Portugal, and the United Kingdom) and two neighbour regions of Catalonia in France (Midi-Pyrenees and Languedoc-Roussillon).

Facebook imposes a lower bound of 1000 users for MAU for privacy reasons, so it reports 1000 instead of lower values. To address this issue, we replace MAU with DAU (daily active users) whenever MAU equals 1000. It should be noted that Facebook's minimum reported DAU is 20. In addition, to avoid the lower bond limitation, we targeted geographic regions with high user counts. We collected data for our analysis by programmatically querying the Facebook Marketing API with 67,271 queries across each geographical region (17 Spanish CCAAs, six European countries and two French regions).

5.1 Representativeness of the data

To determine the representativeness of the Facebook data, we examined the data in relation to Facebook coverage in the studied regions. We obtained the Facebook audience counts by querying the Facebook marketing API. We also obtained the population of each Spanish CCAA from the Spanish National Statistics Institute (www.ine.es). The penetration of Facebook in each region was then calculated by dividing the audience count by the population of each region. The distribution of Facebook penetration in Spanish CCAAs is depicted in Fig. 1. According to the findings, Facebook penetration in all Spanish CCAAs exceeds 50%, with Catalonia and the Community of Madrid having higher penetration than the rest of Spain. We performed a similar analysis for some European countries using population data from the www.statista.com website, and the results are shown in Fig. 2.

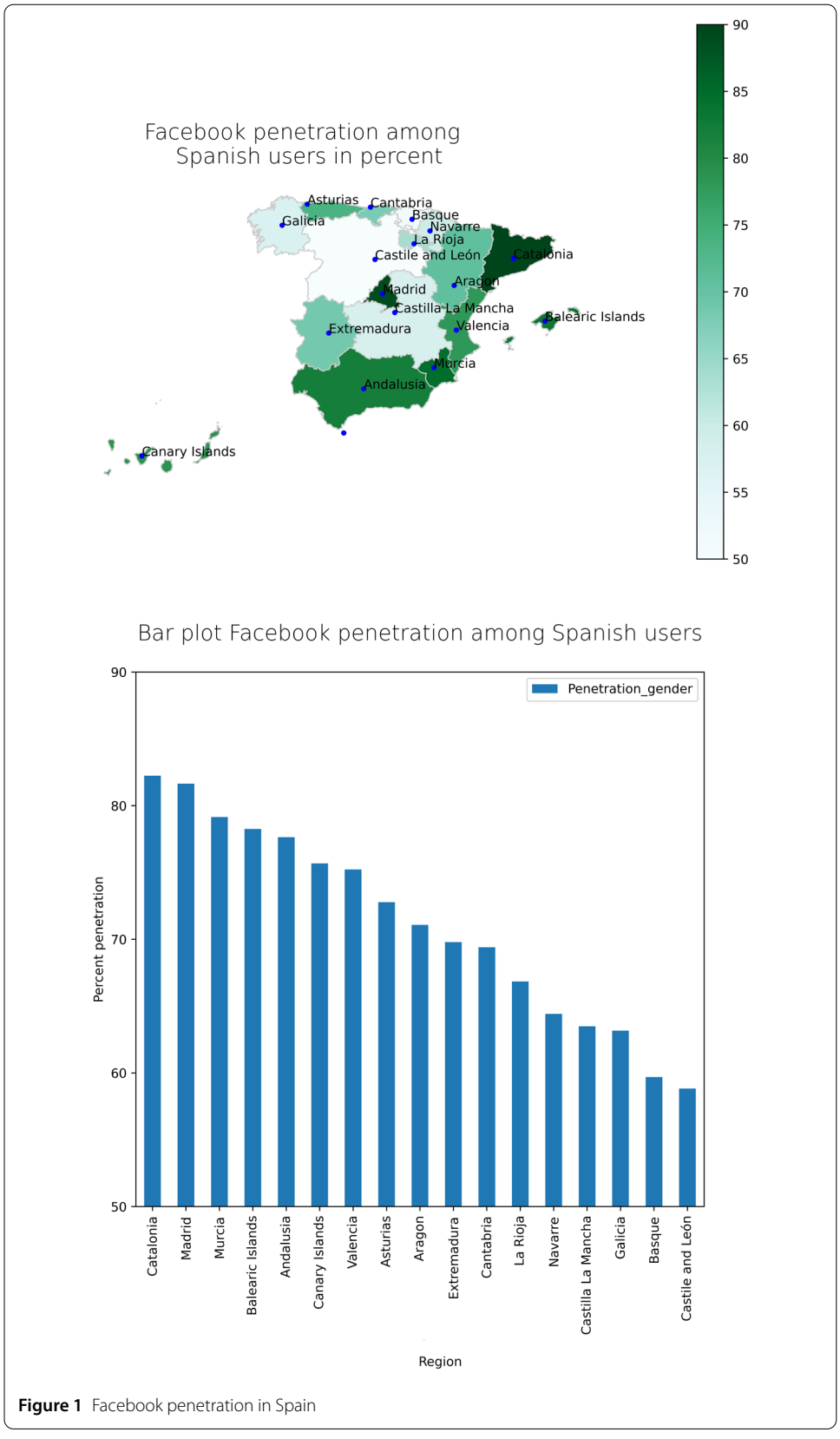

Figure 1 Facebook penetration in Spain

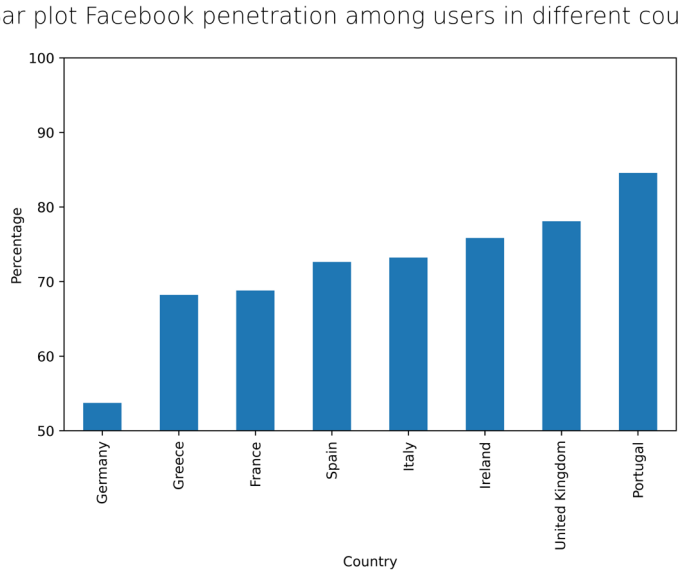

Figure 2 Facebook penetration in euroupian countries

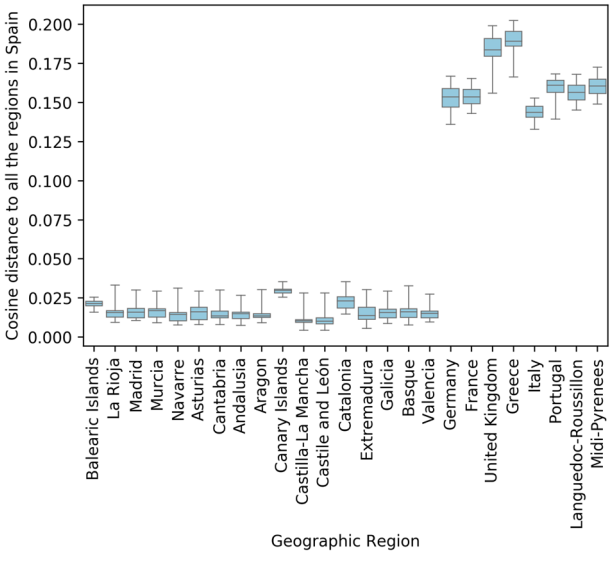

Figure 3 Cosine distance measurement between Spanish CCAAs and different geographic regions

6 Results and discussion

We assumed that the population is divided into 17 subgroups and measured the faultlines separating each CCAA from Spain. This way, we can identify and measure strong faultlines.

6.1 Distance analysis

Figure 3 depicts the cosine distances between the IPVs of Spanish CCAAs and those of several European countries and the considered French regions. Each box plot depicts the distribution of cosine distances between each CCAA's IPV and the remaining Spanish

CCAAs. The boxes for the European countries present the cosine distances between the country/region's IPV and all Spanish CCAAs. The primary observation in Fig. 3 is that the cosine distance between IPVs of the Spanish CCAAs is roughly one order of magnitude lower than the cosine distance between European countries/regions and Spanish CCAAs. Table 2 provides the average distance between the IPVs of geographic regions and Spanish CCAAs. The average distance between IPVs between Spanish CCAAs ranges from 0.01 to 0.03 while the distance between other countries ranges from 0.14 to 0.19. People living in Spanish CCAAs (in-group because Spain is a supergroup for all CCAAs) are an order of magnitude closer compared with people living in other European countries (out-group for people living in Spain). This finding is consistent with our theoretical discussion that members of a group (Spanish CCAAs) feel less distance from one another than members of other groups (Countries).

Moreover, according to Table 2, Spanish Islands (regions with insularity effect) and Catalonia (strong regional identity) present the highest average distance among the Spanish geographic regions. To test these initial observations regarding hypothesis 2 (H2), we used an honestly significant test (HDS) designed by Tukey [46] to compare the distance between distributions pairwise. According to the results of this test, we found enough evidence to reject the null hypothesis for the similarity of Canary Islands to all 16 remaining Spanish CCAAs with 95% certainty (Table 7). This confirms, for Canary Islands, the furthest insular Spanish territory. Similarly, we found enough evidence to reject the null hypothesis for Catalonia's similarity to 11 out of the 16 remaining CCAAs with 95% certainty (Table 8).

Moreover, we compute the HDS for the 17 Spanish CCAAs and the considered 6 European countries and 2 neighboring French provinces. The HDS results allow us to reject the null hypothesis of similarity in all cases. Tables 9, 12, 13, 14, 15, 16, 10 and 11 in the Appendix show the results of this analysis for various regions and countries. When we compare the distances in these tables, we can see that the mean difference between Canary

Table 2 Cosine distance between IPVs of geographic regions and Spanish CCAAs

Geographic Rg	Min	Max	Mean	Median
Balearic Islands	0.016	0.025	0.021	0.021
La Rioja	0.009	0.033	0.016	0.015
Madrid	0.010	0.030	0.016	0.015
Murcia	0.009	0.029	0.016	0.016
Navarre	0.007	0.031	0.014	0.014
Asturias	0.007	0.029	0.016	0.016
Cantabria	0.008	0.030	0.015	0.013
Andalusia	0.007	0.026	0.015	0.015
Aragon	0.009	0.030	0.014	0.013
Canary Islands	0.025	0.035	0.029	0.029
Castilla La Mancha	0.004	0.028	0.011	0.010
Castile and Leon	0.004	0.028	0.012	0.010
Catalonia	0.014	0.035	0.023	0.023
Extremadura	0.005	0.030	0.015	0.013
Galicia	0.008	0.029	0.016	0.015
Basque	0.007	0.032	0.016	0.016
Valencia	0.009	0.027	0.015	0.015
Germany	0.135	0.229	0.158	0.154
France	0.135	0.228	0.160	0.154
United Kingdom	0.098	0.199	0.179	0.183
Greece	0.098	0.208	0.185	0.189
Italy	0.132	0.191	0.148	0.145
Portugal	0.139	0.229	0.171	0.163

Table 3 Cosine distance between regions with strong fultlines and the rest of Spain

CCAA	Canary	Catalonia	Basque	Balearic
Catalonia	0.035	0.001	0.017	0.018
Madrid	0.030	0.014	0.011	0.019
Valencia	0.027	0.016	0.015	0.015
Basque	0.032	0.017	0.001	0.022
Balearic Islands	0.025	0.018	0.022	0.001
Navarre	0.031	0.018	0.007	0.020
Aragon	0.030	0.018	0.012	0.020
Castile and León	0.028	0.021	0.011	0.020
Castilla-La Mancha	0.028	0.022	0.012	0.019
Andalusia	0.026	0.023	0.016	0.019
La Rioja	0.033	0.024	0.012	0.023
Murcia	0.029	0.025	0.017	0.021
Cantabria	0.030	0.025	0.012	0.022
Galicia	0.029	0.026	0.016	0.022
Asturias	0.029	0.028	0.018	0.023
Extremadura	0.030	0.029	0.018	0.023
Canary Islands	0.001	0.035	0.032	0.025

Islands and the rest of the Spanish CCAAs (Table 7) is one order of the magnitude less than the mean difference between other foreign countries/provinces. This finding is aligned well with the first hypothesis (H1). According to Spain's faultline analysis, the strongest faultlines separate Catalonia, Basque, and the Canary Islands from the rest of the regions. Table provides the cosine distance between the Spanish Islands, Catalonia, Basque country to the rest of Spain. We observe that the Canary Islands are far from all other CCAAs compared to distances observed in the other three regions evaluated). Interestingly, the closest region is the Balearic Islands (0.0256), which share the Canary Islands' insularity property. It is worth noting that Canary Islands is the furthest away region from any other Spanish region.

The strongest faultlines separate Catalonia, Basque Country, and the Canary Islands from the rest of Spain, according to faultline analysis. The Table 3 shows the cosine distance from the Spanish Islands, Catalonia, and the Basque Country to the rest of Spain. In comparison to the distances observed in the other three regions evaluated, the Canary Islands are far from all other CCAAs. Surprisingly, the closest region is the Balearic Islands (0.0256), which have the same insularity as the Canary Islands. It is worth noting that when compared to any other region, Canary Island is the farthest away.

The three closest regions to the Balearic Islands are Valencia, Catalonia, and the Community of Madrid, in that order. It is reasonable because the Valencia and Catalonia are the closest geographical regions and share a common local language with the Balearic Islands, and all three have a language that is similar to Catalan. Because Madrid is the country's capital and has strong communication links with all regions, it is reasonable to expect Madrid to never appear among the most distant CCAAs for the other regions.

Analyzing Catalonia, we find six regions with distances less than 0.02: Madrid, the Community of Valencia, the Basque Country, the Balearic Islands, Navarre, and Aragon. Madrid and Catalonia are Spain's most developed and international regions, with good transport links. It allows the people who live in these areas to communicate and exchange cultural values.

Table 3 also shows reciprocity with the Balearic Islands and Valencia due to the common local language and because Valencia is a neighboring region. People who live in border regions develop common cultural values, as is the case in Aragon.

Finally, as we argue in Sect. 3, the Basque Country and Catalonia have strong cultural ties because they both had secessionist political movements. The Basque country secessionist movements have such a strong influence in Navarre that nationalist Basque parties argue that Navarre should be a part of the Basque country. Furthermore, there are many areas in Navarre where some people speak Euskera (the local language spoken in the Basque Country). As a result, Navarre is significantly closer to the Basque country than any other region. The insular regions farthest from the Basque Country are the result of the previously described insular effects.

The Canary Islands are the furthest away from Catalonia, while the Community of Madrid is the closest. This finding is interesting because the two sides of the dispute in this region were the national government based in Madrid and the regional government in Catalonia. Therefore, to make the faultlines in Catalonia less distant, seems to not be enough to manage conflicts between the regional and the national governments.

6.2 Faultline strength

We considered the faultlines that split Spain into two subgroups, i.e., each faultline separates one of the CCAAs from the rest of Spain. Table 4 shows the calculated Fau strength, distance and distance $\times$ strength values for each faultline using our dataset. According to Table 4, the strongest faultlines belong to the Spanish Islands, Catalonia, and the Basque country, which confirms the hypothesis (H2). The existing literature on cultural diversity [43] and the evidence of the historical conflicts in two CCAAs (Catalonia, Basque Autonomous Region) shows that the faultlines in these regions are relatively strong. The measurement results confirm the methodology's validity we introduced in this paper. We analyzed the robustness of our methodology using clustering analysis of the *IPV*s [11]. Clustering relaxes the assumption made by Fau measurement that only two subgroups exist in the group. This clustering analysis aims to find all the subgroups using the IPVs of Spanish CCAAs and report the strongest faultline and the relevant calculated strength. The results confirm the findings by the former measurements by revealing that the strongest

Table 4 Faultline strength (Fau), distance and Bezrukova measurement for Spanish CCAAs

Autonomous Community	Fau	Eclucidian Distance	Bezrukova
Canary Islands	0.169	0.308	1.825
Balearic Islands	0.093	0.126	1.355
Catalonia	0.088	0.116	1.317
Basque	0.088	0.117	1.321
Extremadura	0.077	0.095	1.235
Asturias	0.076	0.093	1.225
Madrid	0.072	0.087	1.197
Galicia	0.055	0.058	1.047
Cantabria	0.050	0.050	0.997
Andalusia	0.049	0.048	0.987
Navarre	0.046	0.043	0.951
La Rioja	0.043	0.040	0.923
Murcia	0.043	0.039	0.922
Aragon	0.034	0.028	0.829
Valencia	0.034	0.028	0.828
Castile and León	0.019	0.012	0.622
Castilla-La Mancha	0.018	0.010	0.596

faultline separates the population in the Canary Islands from all the remaining CCAAs with a faultline strength measurement value equal to 0.29. Figure 4 and Fig. 5 present the distribution of the faultline strength in Spanish geographic regions.

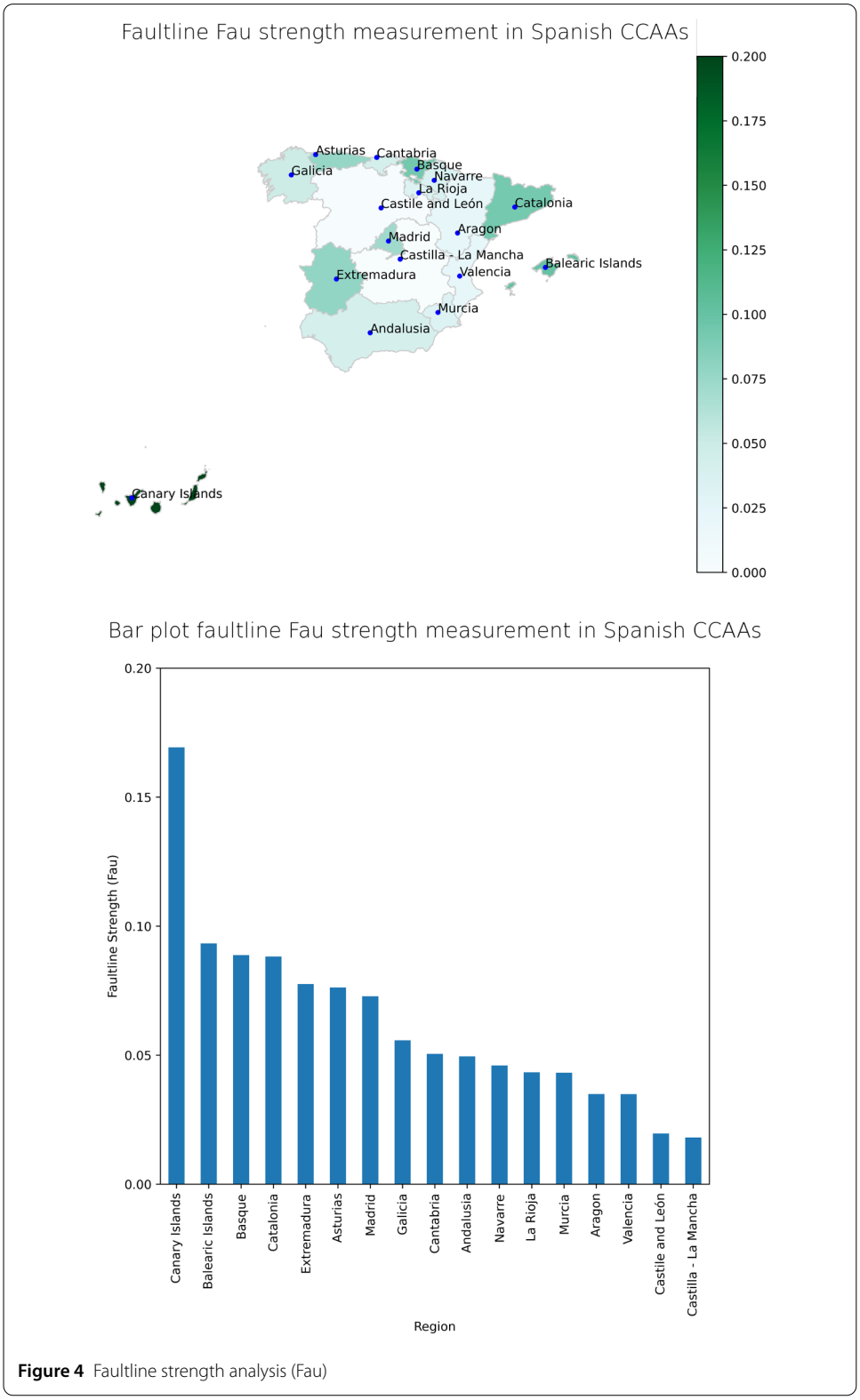

Figure 4 Faultline strength analysis (Fau)

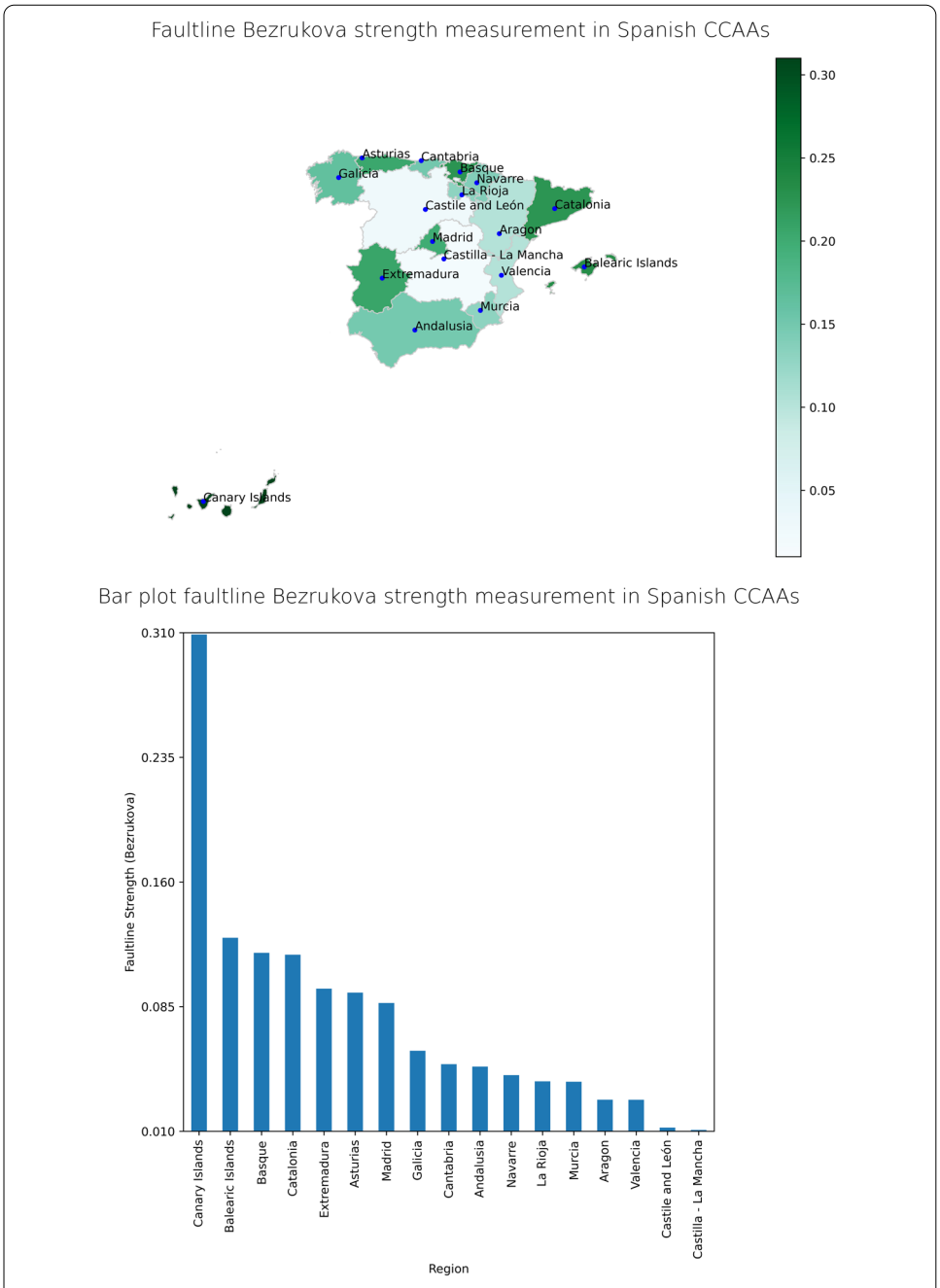

Figure 5 Faultline strength analysis (Bezrukova)

Looking at Fig. 4 we can observe relatively strong faultlines in the Spanish CCAAs. The methodology we introduced in this paper provides politicians with a tool to better allocate resources to manage regional conflicts and reduce similar regional identity-based disputes.

6.3 Singularity analysis

We analyzed interest penetrations in CCAAs to find unique interests in each region. We define *Singularities* as interest categories that satisfy the following conditions: (1) interest

penetration in one region is ten times (10×) larger than in the rest of the regions (2) interest penetration is more than 5% in that region. We repeated our analysis by changing the first conditions' threshold from 10× to 5×. We reported the results of this analysis in Table 5. With a 10× threshold for Catalonia, we found "Time Out Barcelona" and "Catalunya Experience". The first one seems to be corresponding to the political interests of the population in Catalonia. In Rioja, we found "Rioja (wine)" to be the sole singularity using the 10× threshold. Rioja is a touristic region which is famous for the wine [44]. Hence, these results suggest that, using this methodology, we can identify some indentitarian interest categories.

Applying social identity theories to this analysis, we can explain the group-level behavior of the people in Spanish regions. For example, Social identity theory explains that in Rioja, people tend to maintain their self-esteem by a cognitive bias assigning positive attributes such as "the best wine quality" to their regional identity [20]. If the same people belonged to other Spanish regions like "Ribera del Duero," they would probably assume Ribera wine has the best quality.

6.4 Variables which may affect faultline measurement

We did a multiple linear regression to control the economic factors, population, and Facebook penetration in different CCAAs. We normalized the independent variables (Population, GDP and Penetration) and reported the multiple linear regression model coefficients in Table 6 considering the dependent variable here is faultline strength. None of the dependent variables are statistically significant with a 5% threshold, providing strong evidence for the null hypothesis that the dependent variable are not correlated to the independent variables. Therefore, we retain the null hypothesis that there is no relationship between independent and dependent variables and reject the alternative hypothesis.

7 Conclusion

The extensive coverage and tracking capability of online social networks provide new tools for addressing conflict and faultiness research at the scale of large geographical regions, overcoming some limitations of traditional approaches. We contributed to the literature on faultline measurement in this paper by introducing a new methodology for analyzing and measuring faultlines between people living in different geographic regions using data from social network platforms. We first divided people into geographic regions and extracted behavioral attributes from Facebook data. To measure faultlines on an unprecedented scale, we use the well-known faultlines distance and strength measurement techniques. The social science literature has introduced management tools to deal with group faultlines, such as increased communication and cultural diversity. The methodology presented in this paper can help politicians assess the effectiveness of differnet policies in managing these faultlines. Our methodology enables us to: (1) monitor the formation of faultlines by identifying the appearance of strong faultlines (i.e., an increase in faultline distance or strength in specific regions); (2) evaluate the evolution of existing faultlines (e.g., see if the strength or distance of the Catalonia faultline increases or decreases); and (3) have a direct measurement method for the performance of political/social action (e.g., if the Spanish government implements some active policy on the Catalonia conflict, then the impact on the strength/distance of the faultline can be monitored.

We tested our technique's ability to accurately measure faultlines on a large scale using the administrative regions of a large European country like Spain. The obtained results are

Table 5 Table of singularities for each Spanish Region

CCAAs	10×	5×
Balearic Islands	Palma, Majorca, Ibiza (town)	Ibiza, Minorca, Formentera, Yacht, Segunda División, Ibiza (town), Washing machine
La Rioja	Rioja (wine)	La Rioja (Spain), Mayor, Rioja (wine), Wine cellar
Madrid Murcia		Low-cost carrier, Leganés Murcia, Truth, Cartagena, Spain, Cartagena, Colombia, PcComponentes.com, Translation
Navarre	Diário de Notícias	Diário de Notícias, Mexican League
Asturias	Lena Meyer-Landrut, Sporting de Gijón, TeleCable (Spain)	Middle Ages, Cider, Viceroy, IPod, Lena Meyer-Landrut, MP3, Sporting de Gijón, TeleCable (Spain), Firewood, Omelette, Automotive lighting, Queen Letizia of Spain, Cod, David Villa, Radio frequency, Igloo, Independent Online (South Africa), Western world
Cantabria		Santander, Spain, Infinity, Ferry, Racing Club de Avellaneda, Lalín, Playa, Red (Color)
Andalusia		Andalusia, Málaga, Cádiz, Marbella, Province of Málaga, Costa del Sol, Fuengirola, Estepona, Torremolinos, Benalmádena, Mijas, Sherry
Aragon Canary Islands	Santa Cruz de Tenerife, Las Palmas, Fuerteventura, La Palma, Fred (footballer), UD Las Palmas, El Hierro, Rudy (film), Arisa, Corralejo	Tenerife, Atlantic Ocean, Gran Canaria, Santa Cruz de Tenerife, Archipelago, Las Palmas, North Africa, Morocco, Lanzarote, Fuerteventura, La Palma, Media (communication), Fred (footballer), Sand, Maghreb, Santi Cazorla, Santa Cruz, California, UD Las Palmas, Atlantic Records, El Hierro, Fertilizer, Oceans (film), Rudy (film), Tropics, Farmer, Arisa, Corralejo, Hard Rock Cafe, Report
Castilla La Mancha		Toledo, Spain, Cuenca, Ecuador
Castile and Leon	University of Valladolid	Valladolid, University of Valladolid, Zamora, Spain, Ávila, Spain
Catalonia	Time Out Barcelona, Catalunya Experience	Catalonia, Girona, Circuit de Barcelona-Catalunya, Barcelona–El Prat Airport, NacióDigital. Cat, Cheers, Republican Left of Catalonia, Costa Brava, Felis, TV3 (Catalonia), Time Out Barcelona, RAC 1, TV3 (Malaysia), Catalunya Experience, Munich BTV (Bulgaria)
Extremadura	Mérida, Yucatán	Mérida, Yucatán, UNESCO, Lighthouse, Tandoor, Naan, Airbag, Lollipop, Trujillo, Peru
Galicia	Celtic nations	Vigo, Santiago de Compostela, Celtic nations, Santiago, Tempo, Estrella Galicia, Celta de Vigo, River, Nova (TV series), Rio (2011 film), Rio Carnival, Haiti, Region
Basque	Guggenheim Museum Bilbao	Basque Country (autonomous community), Athletic Bilbao, Guggenheim Museum Bilbao, athletic club, Bilbo Baggins, Euskaltel, Eibar
Valencia		Valencia, Valencia CF, Bonaire, University of Alicante, Altea

Table 6 OLS Regression results for correlation between Facebook penetration, GDP and population of each region and the faultline strength (Bezrukova)

Dep. Variable:		Bezrukova		*R-squared:*		0.188
Model:		OLS		*Adj. R-squared:*		0.000
Method:		Least Squares		*F-statistic:*		1.002
				Prob (F-statistic):		0.423
	Coef.	Std. err.	*t*	*P* > |*t*|	[0.025	0.975]
Const.	8.19	5.15	1.59	0.14	−2.94	19.333
Population	−9.86	8.29	−1.19	0.26	−27.7	8.064
GDP	1.65	7.03	0.23	0.82	−13.5	16.839
Penetration	13.8	8.06	1.71	0.11	−3.65	31.185

consistent with the expectations. The administrative regions with the strongest faultlines are the insular CCAAs (the Canary Islands and the Balearic Islands), which are physically separated from the continental Spanish regions, and (ii) the CCAAs with active political conflicts in recent Spanish history (Catalonia and Basque Country), which developed stronger regional identities. The literature in social sciences emphasizes the importance of identifying faultlines as a source of conflict in order to take corrective measures. In this regard, our methodology contributes to the field by bridging the gap between faultlines theory and its practical application in large-scale contexts. While we studied faultlines that separate groups of people divided by geographic regions in this paper; faultlines between generations or gender can also be analyzed with the proposed methodology.

Finally, we would like to expand our work to other countries or collaborate with multidisciplinary researchers who are willing to do so in the future. We decided to focus our efforts on only one country due to the technical complexity, the time required to collect the data, and the requirement of explaining the context of the country under analysis to allow the reader to assess the faultline strength results (see Sect. 3).

7.1 Limitations

On the one hand, This research is based on data collected from online social networking platforms. The underlying assumption in our analysis is that social network users, specifically Facebook users, properly represent the entire population in the considered geographic regions. Despite the fact that social network users represent a significant portion of the population in each considered geographic region, certain socio-demographic groups may be underrepresented in the Facebook user database. In our study, we consider the entire population of a region as our study group. Facebook's coverage is over 50% of the population in every region. Hence, we can assert that we have a coverage of the representativeness of each of the 67,270 behavioral interest topics across at least 50% of the considered population. We believe that despite its imperfection our methodology offer better representativeness than traditional methods in the considered large-scale usecase.

On the other hand, this paper does not study how each of the considered interests contribute to the distance or the strenght of faultlines. This requires a very involved and deep analysis that we leave for future research. To this end we plan to use different machine learning techniques (e.g., regression models) as well as other more traditional techniques such as Principal Components Analysis (PCA).

Appendix

This Appendix present the results of the HSD Tukey test comparing each of the 6 European countries considered and the 2 French regions. The results are presented in Tables 7, 8, 10, 11, 9, 12, 13, 14, 15 and 16.

Table 7 HSD Tukey test filtered by for the pair of regions which include Catalonia and we found enough evidence to support our hypothesis with 95% accuracy

Pair of Catalonia and Spanish Regions	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Andalusia	0.007	0.001	0.014	6e–3
Aragon	0.008	0.001	0.015	2e–3
Basque	0.006	0.001	0.013	4e–2
Canary Islands	0.006	0.013	0.000	3e–2
Cantabria	0.007	0.001	0.014	1e–2
Castile and León	0.010	0.004	0.017	1e–4
Castilla La Mancha	0.011	0.004	0.017	1e–4
Madrid	0.006	0.013	0.000	4e–2
Valencia	0.007	0.014	0.001	6e–3
Extremadura	0.007	0.014	0.000	1e–2
Navarre	0.008	0.015	0.001	2e–3

Table 8 HSD Tukey test filtered by for the pair of regions which include Canary Islands and we found enough evidence to support our hypothesis with 95% accuracy

Pair of Canary Islands and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Andalusia	0.014	0.008	0.021	4e–11
Aragon	0.015	0.008	0.022	1e–11
Balearic Islands	0.008	0.001	0.015	1e–03
Basque	0.013	0.007	0.020	1e–09
Cantabria	0.014	0.021	0.007	2e–10
Castile and León	0.017	0.024	0.011	3e–13
Castilla-La Mancha	0.017	0.024	0.011	3e–13
Catalonia	0.006	0.013	0.001	3e–02
Madrid	0.013	0.020	0.007	1e–09
Valencia	0.014	0.021	0.008	4e–11
Extremadura	0.014	0.021	0.007	1e–10
Galicia	0.013	0.020	0.006	3e–09
La Rioja	0.013	0.020	0.006	2e–09
Navarre	0.015	0.022	0.008	9e–12
Asturias	0.013	0.019	0.006	9e–09
Murcia	0.013	0.019	0.006	8e–09

Table 9 HSD Tukey test results filtered by for the pair of CCAAs-countires which include France

Pair of France and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Spain	0.091	0.155	0.027	1e–4
Navarre	0.087	0.151	0.022	2e–4
Murcia	0.084	0.148	0.020	4e–4
La Rioja	0.084	0.148	0.020	4e–4
Galicia	0.082	0.146	0.018	6e–4
Asturias	0.082	0.146	0.018	7e–4
Canary Islands	0.075	0.011	0.139	8e–4
Extremadura	0.081	0.017	0.145	9e–4
Cantabria	0.083	0.019	0.147	5e–4
Balearic Islands	0.084	0.020	0.148	4e–4
Catalonia	0.084	0.020	0.149	4e–4
Andalusia	0.085	0.021	0.149	3e–4
Basque	0.086	0.022	0.150	3e–4
Castile and León	0.086	0.022	0.150	2e–4
Aragon	0.086	0.022	0.150	2e–4
Castilla La Mancha	0.086	0.022	0.150	2e–4
Madrid	0.087	0.023	0.151	2e–4
Valencia	0.087	0.023	0.151	2e–4

Table 10 HSD Tukey test results filtered by for the pair of CCAAs-countires which include Midi-Pyrenees

Pair of close French province (Midi-Pyrenees) and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Andalusia	0.092	0.028	0.156	4e–5
Aragon	0.093	0.028	0.157	3e–5
Balearic Islands	0.091	0.027	0.155	5e–5
Basque	0.092	0.028	0.156	3e–5
Canary Islands	0.082	0.018	0.146	7e–4
Cantabria	0.090	0.026	0.154	7e–5
Castile and León	0.093	0.028	0.157	3e–5
Castilla La Mancha	0.093	0.029	0.157	3e–5
Catalonia	0.091	0.027	0.155	5e–5
Madrid	0.094	0.029	0.157	2e–5
Valencia	0.094	0.030	0.158	2e–5
Extremadura	0.088	0.024	0.152	1e–4
Galicia	0.090	0.025	0.153	1e–4
La Rioja	0.091	0.027	0.155	6e–5
Navarre	0.094	0.157	0.029	3e–5
Asturias	0.089	0.153	0.025	1e–4
Murcia	0.091	0.155	0.027	6e–5
Spain	0.099	0.162	0.034	1e–5

Table 11 HSD Tukey test results filtered by for the pair of CCAAs-countires which include Languedoc-Roussillon

Pair of close French province (Languedoc-Roussillon) and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Andalusia	0.088	0.024	0.152	1e–4
Aragon	0.089	0.024	0.153	1e–4
Balearic Islands	0.087	0.023	0.151	1e–4
Basque	0.088	0.024	0.152	1e–4
Canary Islands	0.078	0.014	0.142	2e–3
Cantabria	0.086	0.022	0.150	2e–4
Castile and León	0.088	0.024	0.153	1e–4
Castilla-La Mancha	0.089	0.025	0.153	1e–4
Catalonia	0.087	0.023	0.151	1e–4
Madrid	0.089	0.025	0.153	0e–5
Valencia	0.090	0.026	0.154	0e–5
Extremadura	0.084	0.020	0.148	4e–4
Galicia	0.085	0.021	0.149	3e–4
La Rioja	0.087	0.023	0.151	2e–4
Navarre	0.089	0.153	0.025	9e–5
Asturias	0.085	0.149	0.021	3e–4
Murcia	0.087	0.151	0.023	1e–4
Spain	0.094	0.158	0.030	2e–5

Table 12 HSD Tukey test results filtered by for the pair of CCAAs-countires which include United Kingdom

Pair of France and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Spain	0.091	0.155	0.027	5e–5
Navarre	0.086	0.151	0.022	2e–4
Murcia	0.084	0.148	0.020	3e–4
La Rioja	0.084	0.148	0.020	4e–4
Galicia	0.082	0.146	0.018	6e–4
Asturias	0.082	0.146	0.018	6e–4
Canary Islands	0.075	0.011	0.139	4e–3
Extremadura	0.081	0.017	0.145	8e–4
Cantabria	0.083	0.019	0.147	5e–4
Balearic Islands	0.084	0.020	0.148	3e–4
Catalonia	0.084	0.020	0.148	3e–4
Andalusia	0.085	0.021	0.149	2e–4
Basque	0.086	0.022	0.150	2e–4
Castile and León	0.086	0.022	0.150	2e–4
Aragon	0.086	0.022	0.150	2e–4
Castilla La Mancha	0.086	0.022	0.150	2e–4
Madrid	0.087	0.023	0.151	1e–4
Valencia	0.087	0.023	0.151	1e–4

Table 13 HSD Tukey test results filtered by for the pair of CCAAs-countires which include Germany

Pair of Germany and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Andalusia	0.094	0.030	0.158	2e–5
Aragon	0.094	0.030	0.158	2e–5
Balearic Islands	0.093	0.029	0.157	3e–5
Basque	0.094	0.030	0.158	2e–5
Canary Islands	0.084	0.020	0.148	4e–4
Cantabria	0.092	0.028	0.156	4e–5
Castile and León	0.094	0.030	0.158	2e–5
Castilla-La Mancha	0.094	0.030	0.158	2e–5
Catalonia	0.093	0.029	0.157	3e–5
Madrid	0.095	0.031	0.159	1e–5
Valencia	0.096	0.032	0.160	1e–5
Extremadura	0.090	0.026	0.154	8e–5
Galicia	0.091	0.027	0.155	5e–5
La Rioja	0.092	0.156	0.028	3e–5
Navarre	0.095	0.159	0.031	1e–5
Asturias	0.091	0.155	0.027	6e–5
Murcia	0.093	0.157	0.029	3e–5
Spain	0.100	0.164	0.036	4e–7

Table 14 HSD Tukey test results filtered by for the pair of CCAAs-countires which include Italy

Pair of Italy and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Andalusia	0.086	0.022	0.150	2e–4
Aragon	0.087	0.023	0.151	2e–4
Balearic Islands	0.085	0.021	0.149	3e–4
Basque	0.086	0.022	0.150	2e–4
Canary Islands	0.076	0.012	0.140	3e–3
Cantabria	0.084	0.020	0.148	4e–4
Castile and León	0.087	0.022	0.151	2e–4
Castilla-La Mancha	0.087	0.023	0.151	2e–4
Catalonia	0.085	0.021	0.149	3e–4
Madrid	0.087	0.023	0.151	1e–4
Valencia	0.088	0.024	0.152	1e–4
Extremadura	0.082	0.018	0.146	7e–4
Galicia	0.083	0.019	0.147	5e–4
La Rioja	0.085	0.149	0.021	3e–4
Navarre	0.087	0.151	0.023	1e–4
Asturias	0.083	0.147	0.019	5e–4
Murcia	0.085	0.149	0.021	3e–4
Spain	0.092	0.156	0.028	4e–5

Table 15 HSD Tukey test results filtered by for the pair of CCAAs-countires which include Greece

Pair of Greece and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Andalusia	0.125	0.061	0.189	5e–10
Aragon	0.126	0.062	0.190	4e–10
Balearic Islands	0.124	0.060	0.188	7e–10
Basque	0.126	0.062	0.190	5e–10
Canary Islands	0.115	0.051	0.179	1e–8
Cantabria	0.123	0.059	0.187	1e–9
Castile and León	0.126	0.062	0.190	4e–10
Castilla-La Mancha	0.126	0.062	0.190	4e–10
Catalonia	0.124	0.060	0.188	7e–10
Madrid	0.127	0.063	0.191	3e–10
Valencia	0.127	0.063	0.191	3e–10
Extremadura	0.121	0.057	0.185	2e–9
Galicia	0.122	0.058	0.186	1e–9
La Rioja	0.124	0.188	0.060	8e–10
Navarre	0.126	0.191	0.062	4e–10
Asturias	0.122	0.186	0.058	1e–9
Murcia	0.124	0.188	0.060	7e–10
Spain	0.131	0.195	0.067	2e–10

Table 16 HSD Tukey test results filtered by for the pair of CCAAs-countires which include Portugal

Pair of Portugal and Spanish CCAAs	Mean Difference	95% Confidence interval		Adjusted P-value
		Lower band	Upper bound	
Andalusia	0.113	0.049	0.177	4e–8
Aragon	0.113	0.049	0.177	3e–8
Balearic Islands	0.112	0.048	0.176	5e–8
Basque	0.113	0.049	0.177	3e–8
Canary Islands	0.103	0.039	0.167	1e–6
Cantabria	0.111	0.047	0.175	8e–8
Castile and León	0.113	0.049	0.177	3e–8
Castilla La Mancha	0.113	0.049	0.177	3e–8
Catalonia	0.112	0.048	0.176	5e–8
Madrid	0.114	0.050	0.178	2e–8
Valencia	0.115	0.051	0.179	2e–8
Extremadura	0.109	0.045	0.173	1e–7
Galicia	0.110	0.046	0.174	1e–7
La Rioja	0.111	0.047	0.175	6e–8
Navarre	0.114	0.050	0.178	2e–8
Asturias	0.110	0.174	0.046	1e–7
Murcia	0.112	0.176	0.048	6e–8
Spain	0.119	0.183	0.055	4e–9

Acknowledgements
We would like to thank the editor and reviewers for their insightful comments, which helped improved the quality of this paper.

Funding
This research received funding from the European Union's Horizon 2020 innovation action program under the PIMCITY (Grant 871370) and TESTABLE (101019206) projects. The Agencia Estatal de Investigación (AEI) under the ACHILLES project (Grant PID2019-104207RB-I00/AEI/10.13039/501100011033). The Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M ("Fostering Young Doctors Research"). Finally, the Fundación BBVA under the project AERIS.

Abbreviations
CCAA, comunidades autónomas; FDVT, Facebook data valuation tool; MPAIAC, movimiento por la autodeterminación e independencia del Archipiélago Canario; ETA, Euskadi 'ta Askatasuna; MLNV, Basque national liberation movement; API, application programming interface; WVS, world values survey; MAU, monthly active users; DAU, daily active users; IPV, interest penetration vector; IP, interest penetration; KAS, Koordinadora Abertzale Sozialista.

Availability of data and materials
The data and the source code we developed to do this research is publicly available on the following online repository.[4]

Declarations

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
RC contributed to the design of the paper, the development of the measurement methodology, and the paper writing. ARM contributed to the design of the paper, the execution of the experiments, and the paper writing. AC contributed to the design of the paper and the development of the measurement methodology, and the paper writing. All authors read and approved the final manuscript.

Author details
[1]IMDEA Networks Institute, Leganes, Spain. [2]Telematics Engineering Department, Universidad Carlos III of Madrid (UC3M), Madrid, Spain. [3]Big Data Institute, UC3M-Santander, Getafe, Spain.

Publisher's Note

Received: 25 November 2021 Accepted: 13 June 2022 Published online: 07 July 2022

References

1. Bowers CA, Pharmer JA, Salas E (2000) When member homogeneity is needed in work teams: a meta-analysis. Small Group Res 31(3):305–327
2. Webber SS, Donahue LM (2001) Impact of highly and less job-related diversity on work group cohesion and performance: a meta-analysis. J Manag 27(2):141–162
3. Horwitz SK, Horwitz IB (2007) The effects of team diversity on team outcomes: a meta-analytic review of team demography. J Manag 33(6):987–1015
4. Bell ST, Villado AJ, Lukasik MA, Belau L, Briggs AL (2011) Getting specific about demographic diversity variable and team performance relationships: a meta-analysis. J Manag 37(3):709–743
5. Guillaume YRF, Brodbeck FC, Riketta M (2012) Surface-and deep-level dissimilarity effects on social integration and individual effectiveness related outcomes in work groups: a meta-analytic integration. J Occup Organ Psychol 85(1):80–115
6. Van Dijk H, Van Engen ML, Van Knippenberg D (2012) Defying conventional wisdom: a meta-analytical examination of the differences between demographic and job-related diversity relationships with performance. Organ Behav Hum Decis Process 119(1):38–53
7. Lau DC, Murnighan JK (1998) Demographic diversity and faultlines: the compositional dynamics of organizational groups. Acad Manag Rev 23(2):325–340
8. Thatcher S, Patel PC (2011) Demographic faultlines: a meta-analysis of the literature. J Appl Psychol 96(6):1119–1139
9. Rico R, Sánchez-Manzanares M, Antino M, Lau D (2012) Bridging team faultlines by combining task role assignment and goal structure strategies. J Appl Psychol 97(2):407–420
10. Thatcher SMB, Patel PC (2012) Group faultlines: a review, integration, and guide to future research. J Manag 38(4):969–1009
11. Meyer B, Glenz A (2013) Team faultline measures: a computational comparison and a new approach to multiple subgroups. Organ Res Methods 16(3):393–424
12. Meyer B, Glenz A, Antino M, Rico R, González-Romá V (2014) Faultlines and subgroups: a meta-review and measurement guide. Small Group Res 45(6):633–670
13. Lau DC, Murnighan JK (2005) Interactions within groups and subgroups: the effects of demographic faultlines. Acad Manag J 48(4):645–659
14. Turner JC, Hogg MA, Oakes PJ, Reicher SD, Wetherell MS (1987) Rediscovering the social group: a self-categorization theory. Blackwell Sci., Oxford
15. Choi JN, Sy T (2010) Group-level organizational citizenship behavior: effects of demographic faultlines and conflict in small work groups. J Organ Behav 31(7):1032–1054
16. Carton AM, Cummings JN (2012) A theory of subgroups in work teams. Acad Manag Rev 37(3):441–470
17. Brewer MB, Yuki M (2007) Culture and social identity
18. Shivhare R, Kumar CA (2016) On the cognitive process of abstraction. Proc Comput Sci 89:243–252
19. Postmes T, Haslam SA, Swaab RI (2005) Social influence in small groups: an interactive model of social identity formation. Eur Rev Soc Psychol 16(1):1–42
20. Hatch MJ, Schultz M (2004) Organizational identity: a reader. Oxford University Press, Oxford
21. Byrne DE (1971) The attraction paradigm, vol 462. Academic Press, New York
22. Brewer MB (1991) The social self: on being the same and different at the same time. Pers Soc Psychol Bull 17(5):475–482
23. Hais SC, Hogg MA, Duck JM (1997) Self-categorization and leadership: effects of group prototypicality and leader stereotypicality. Pers Soc Psychol Bull 23(10):1087–1099

[4]github.com/AmirXmj/Faultline_calculator

24. Fitzhugh B, Hunt TL (1997) Introduction: islands as laboratories: archaeological research in comparative perspective. Hum Ecol 25(3):379–383
25. Pungetti G (2012) Islands, culture, landscape and seascape. J Mar Isl Cultures 1(2):51–54
26. Jehn KA, Bezrukova K, Thatcher S (2008). Conflict, diversity, and faultlines in workgroups
27. Van der Kamp M, Tjemkes BV, Jehn KA (2012) The rise and fall of subgroups and conflict in teams: faultline activation and deactivation. In: Intl. Association for conflict management, IACM 25th annual conference
28. Li J, Hambrick DC (2005) Factional groups: a new vantage on demographic faultlines, conflict, and disintegration in work teams. Acad Manag J 48(5):794–813
29. Polzer JT, Crisp CB, Jarvenpaa SL, Kim JW (2006) Extending the faultline model to geographically dispersed teams: how colocated subgroups can impair group functioning. Acad Manag J 49(4):679–692
30. Tafarodi RW, Kang S-J, Milne AB (2002) When different becomes similar: compensatory conformity in bicultural visible minorities. Pers Soc Psychol Bull 28(8):1131–1142
31. Van Baaren RB, Holland RW, Kawakami K, Van Knippenberg A (2004) Mimicry and prosocial behavior. Psychol Sci 15(1):71–74
32. Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M (2009) Life in the network: the coming age of computational social science, vol 323. Science, New York, p 721
33. Murthy D (2008) Digital ethnography: an examination of the use of new technologies for social research. Sociology 42(5):837–855
34. Kosinski M, Matz SC, Gosling SD, Popov V, Stillwell D (2015) Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. Am Psychol 70(6):543–556
35. Englehardt S, Narayanan A (2016) Online tracking: a 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 1388–1401
36. Cabañas JG, Cuevas Á, Cuevas R (2018) Unveiling and quantifying Facebook exploitation of sensitive personal data for advertising purposes. In: 27th USENIX security symposium (USENIX security 18), pp 479–495
37. Obradovich N, Özak Ö, Martín I, Ortuño-Ortín I, Awad E, Cebrián M, Cuevas R, Desmet K, Rahwan I, Cuevas Á (2020) Expanding the measurement of culture with a sample of two billion humans. Technical report
38. Bezrukova K, Jehn KA, Zanutto EL, Thatcher SMB (2009) Do workgroup faultlines help or hurt? A moderated model of faultlines, team identification, and group performance. Organ Sci 20(1):35–50
39. Thatcher SMB, Jehn KA, Zanutto E (2003) Cracks in diversity research: the effects of diversity faultlines on conflict and performance. Group Decis Negot 12(3):217–241
40. Esteban J, Schneider G (2008) Polarization and conflict: theoretical and empirical issues. Sage, London
41. Whitfield T (2014) Endgame for ETA: elusive peace in the Basque country. Oxford University Press, Oxford
42. de Felipe MB (2019) Myths and realities of secessionisms: a constitutional approach to the catalonian crisis. Springer, New York
43. Lecours A (2001) Regionalism, cultural diversity and the state in Spain. J Multiling Multicult Dev 22(3):210–226
44. Moreno L (2008) La Federalización de España. Poder Político y territorio Siglo XXI de España Editores, Mexico City, pp 131–132
45. González Cabañas J, Cuevas A, Cuevas R (2017) FDVT: data valuation tool for Facebook users. In: Proceedings of the 2017 CHI conference on human factors in computing systems, pp 3799–3809
46. Abdi H, Williams LJ (2010) Tukey's honestly significant difference (hsd) test. Encycl Res Des 3(1):1–5