Mapping social conflicts in natural resources: a text mining study of extractive activities

Ramiro Albrieu and Gabriel Palazzo

Abstract

Applying text mining techniques, a methodology was developed to measure the number of social conflicts related to the exploitation of non-renewable natural resources. The study focuses on conflicts in four mining countries (Australia, Canada, Chile and Peru) between 2003 and 2016, based on more than 20,000 articles from the leading newspapers of each country. A statically significant correlation was found between the main index and mineral rents as a percentage of gross domestic product (GDP). However, the results should be interpreted with caution since endogeneity issues have not been addressed and the indices could be biased by various, country-specific factors. This study's main outcome is a database with different indices of soft conflicts related to the exploitation of non-renewables natural resources and its media coverage in Australia, Canada, Chile and Peru.

Keywords

Social conflict, natural resources, non-renewable resources, mining, statistical data, Australia, Canada, Chile, Peru

JEL classification

J23, J24, O33

Authors

Ramiro Albrieu is an associate researcher in the Economics Department of the Center for the Study of State and Society (CEDES). Email: ralbrieu@cedes.org.

Gabriel Palazzo is a research assistant in the Economics Department of the Center for the Study of State and Society (CEDES). Email: gabrielmpalazzo@gmail.com.

I. Introduction

The exploitation of underground natural resources is a controversial issue. On the one hand, it can boost government revenues and provide the economy with the necessary income for growth, what Hirschman (1977) called indirect linkages. On the other hand, there is a perception, both in the literature and among the general public, that the social costs of these exploitation activities are not given due consideration when governments (or companies) decide to deplete a given non-renewable natural resource. This is particularly true when it comes to the effects of current actions on the well-being of future generations, but also to their contemporary local effects.

The climate change agenda and, more specifically, the literature on green accounting and its main applications (such as the System of Environmental-Economic Accounting (SEEA) developed by the United Nations) seeks to address the issue of intergenerational equity (United Nations, n/d). The World Bank (2006 and 2011) estimates of adjusted net saving tackle the intergenerational problem by correcting standard accounting savings (the sum of present and future well-being) and measuring investments in human and physical capital, the depletion of natural resources, and environmental damage caused by carbon dioxide and other emissions. However, the issue of intragenerational equity has been harder to address.

This article seeks to contribute to the literature on the exploitation and depletion of natural resources by finding proxy measures for social conflicts at the national and local (or regional) levels. The main goal is to draw attention to conflicts related to the exploitation of natural resources that national institutions and the private sector must address. To that end, patterns of words related to social conflicts were identified in articles about the mining sector published in the leading newspapers of four countries that are major producers of minerals, namely Australia, Canada, Chile and Peru. The main contribution of this study has been the creation of a database to explore causal effects and correlation in an effort to ensure better conflict management in the future.

The advantage of these indices is that they are able to capture minor conflicts and to measure the intensity of different conflicts. But a major weakness is that they may contain biases as a result of the different degrees of lobbying by the extractive industries, government or other agents on the media. However, if the impact of lobbying remained stable over time, the indices in the country sample will only be biased at the country level. This could be easily resolved by using fixed-effect regression models. Problems arise if lobbying varied over time or was subject to temporal shocks. Nevertheless, even if there is some bias, it is still interesting to evaluate the media coverage of social conflicts related to natural resources.

This article is organized as follows. Section II contains a brief overview of the literature on text mining in economics and explains the methodology applied to measure social conflict. Section III presents the main results at the country level, addresses regional disparities within countries, and defines the different levels of violence in the conflict indices measured. Section IV contains the regression models used to evaluate the relationship between conflicts and a country's mineral income and overall economic performance. Lastly, the conclusions are set out in section V.

II. Quantifying social conflicts in extractive industries

1. Literature review

The methodology used in this study is rooted in text mining techniques. These techniques allow conclusions to be reached, foster computational research and detect statistical patterns by studying the words present in a text.

Online newspapers contain vast amounts of qualitative information that can be processed using new software to obtain quantitative assessments for hard-to-quantify economic variables. Gupta and Lehal (2009) describe text mining as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" (p. 60).

The literature on the use of these techniques in economics has grown over time. Tetlock (2007) is one of the pioneers in this area, constructing a measure of media pessimism, mining the daily "Abreast of the Market" column in the Wall Street Journal. Using basic vector autoregressions (VARs), he finds that high values of media pessimism robustly predict downward pressure on market prices, followed by a reversion to fundamentals. Tetlock's findings suggest that measures of media content serve as a proxy for investor sentiment. By the same token, García (2013) constructs an index of market sentiment by counting the number of positive and negative words of two financial columns ("Financial Markets" and "Topics in Wall Street") from the New York Times, which were both published daily over the period 1905–2005. In accordance with Tetlock (2007), García finds that media content can predict trading volume. Using a parsimonious time-series model, García also finds that news content helps to predict daily stock returns, particularly during recessions. Aromí (2013) applies a similar methodology to evaluate how information flows from newspapers influenced the performance of the financial market in Argentina from October 1996 to December 2012. He carries out time-series regression models on stock returns, the determinants of which are media sentiment measures that he has constructed, the lags of these variables and the stock returns lags, among other control variables. Aromí finds evidence compatible with the presence of market participants who overreact to information flows. To obtain a quantitative index all these studies used a dictionary approach, counting positive and negative words to generate a numerical index.

Baker, Bloom and Davis (2013) use text mining techniques to construct an index that quantifies the uncertainty surrounding economic policy and its impact on investment demand in the context of the 2007–2009 economic recession in the United States. They built an economic policy uncertainty (EPU) index which sought to capture the prevailing uncertainty about tax, spending, regulatory and monetary policies. The EPU index is based on an automated text search for terms related to economic policy uncertainty in 10 leading newspapers published in the United States.

Social conflicts over natural resources have a much longer history than the text mining methodology used in this paper. Since the 1980s, the conventional view that natural resource endowments promote economic development has been called into question, with many contending that those very endowments are at the root of underdevelopment. Sachs and Warner (1995) attribute the lower growth exhibited by countries with a high share of commodity exports for the period 1970–1989 to the curse of natural resources. More recently, Brunnschweiler and Bulte (2009) state that there are no less than three different dimensions of the curse of natural resources: (i) slower economic growth; (ii) violent civil conflict; and (iii) undemocratic regimes.

Several reasons may link the exploitation of natural resources to social unrest and political conflicts. Possible hypotheses put forward by Sachs and Warner (1995) and Leite and Weidmann (1999) are that natural resource-rich countries exhibit higher inequality and social polarization because the ruling elite takes advantage of its political power to lobby for resources, fostering rent-seeking behaviour. Collier and Hoeffler (1998, 2004 and 2005) analyse different causes of civil wars and rebellions, among which an abundance of natural resources seems to have a considerable effect. Collier and Hoeffler (2004) contrast the hypothesis that rebellions may be explained by atypically severe grievances with the idea that they are caused by atypical opportunities for building a rebel organization. Their results support the opportunity hypothesis. They interpret the positive relationship between primary commodity exports and higher conflict risk as being due to the opportunities such commodities provide for extortion, making rebellion feasible and perhaps even attractive.

Along a similar line, Nafziger and Auvinen (2002) and Sinnott, Nash and De la Torre (2010) suggest that conflicts could be arise because different social groups do not receive what they consider to be their fair share, indicating a predatory State, weak regulations and elites that leverage resources to extract rents rather than promote economic growth. Sinnott, Nash and De la Torre (2010) identify the social consequences of mining and oil exploitation, which have high potential to generate social tensions and conflicts, in many cases because of their adverse environmental impact and poor working conditions. In turn, the "Dutch Disease" theory and the Prebisch-Singer hypothesis offer explanations as to why the exploitation of natural resources and its impact on the economy can generate social unrest among certain segments of the population (Singer, 1950; Prebisch, 1950).

Brunnschweiler and Bulte (2009) argue that it is necessary to instrument the variables involved. The authors find that there is endogeneity between these variables and reverse causality, whereby peace reduces a country's dependence on natural resources and that it is not the dependency on commodity exports that creates social conflict, which runs contrary to the resource curse thesis; in fact, they find that an abundance of resources could have a positive impact on economic growth. Mehlum, Moene and Torvik (2006), like Brunnschweiler and Bulte (2009), Arezki and van der Ploeg (2007), Haber and Menaldo (2011), and Leite and Wiedmann (1999), claim that the quality of institutions determines whether natural resource abundance is a blessing or a curse. Meanwhile, Giordano, Giordano and Wolf (2005) and Evans (2010) posit that natural resource scarcity may increase the risk of future sociopolitical conflicts. Robinson, Torvik and Verdier (2006) built a political economy model to analyse the impact of natural resources on development. They propose a model where politicians permit the over extraction of natural resources and then engage in inefficient redistribution of the resulting revenues to try to influence elections. However, they conclude that in countries with institutions which limit the ability of politicians to use clientelism to bias elections, resource booms tend to raise national income.

At the time of writing, few studies have used text mining to examine the relationship between natural resource exploitation and sociopolitical conflicts. To our knowledge, the methodology adopted by UNDP/Fundación UNIR (2012) is the closest to ours since they also use media coverage to analyse conflicts. For that study of social conflict in Latin America, data were collected from 54 newspapers published in 17 Latin American countries from October 2009 to September 2010. The authors identify three different spheres of conflict: (i) social reproduction (which account for the largest share of conflicts in the period analysed and includes those related to work/wages, land and incomes); (ii) institutional conflicts (related to demands for practical improvements in the provision of public goods, administrative management and the legitimacy of public authorities); and (iii) cultural conflicts (related to ideological-political issues, public safety and the environment, among other things).

Another example in this field is Dube and Vargas (2013), who use newspaper articles to characterize violent civil conflicts in Colombia. They present evidence that a rise in the price of oil intensified violence in areas transporting and producing more oil, while a fall in the price of coffee increased violence in municipalities growing more coffee. The key difference between these activities is the labour intensity gap. On one hand, the fall in coffee prices reduces workers' wages and lowers the cost of recruiting workers into armed groups. This is called opportunity cost effect. On the other hand, the oil shock substantially increases local government revenue, encouraging paramilitary groups to move into oil areas to control these resources. This is called the rapacity effect. Unlike the present study, however, Dube and Vargas use a dataset from the Conflict Analysis Resource Center (CERAC) collected from newspaper articles, rather than constructing a new database using text mining technique.

Lastly, the empirical strategy used in this paper is borrowed from Palazzo (2017), who created social conflict indices related to the exploitation of a broad set of natural resources in Argentina over the period 1996–2014. This was a major contribution to the development of the methodology and procedure for verifying whether those indices reliably showed some stylized facts about civil conflicts related to agriculture, mining, oil, fishing and forestry activities in Argentina.

2. Empirical strategy used

We have developed a methodology that measures the quantity of sociopolitical conflicts related to the exploitation of mineral resources, that might be applied to other natural resources (see Palazzo, 2017). A text mining bag-of-words model was used, which consisted of counting the number of hostile words in each article that referenced the extractive industries in a particular location and time period. The number of articles about conflicts serves as a proxy for the level of conflict at a particular point in time and in a particular place. We then used this information to create a set of indices (on, for example, the ratio of hostile words to total words per article) to capture the intensity of conflicts.

The majority of the literature on social conflicts examines civil wars where the number of dead people is the measure of intensity. Our index differs from those constructed by other authors, by offering a soft measure of conflict that takes into account strikes, lockouts, protest marches and political struggles. In addition, the conflicts are unequivocally related to the extractive industries because of the nature of the methodology.

The data was taken from Factiva (Dow Jones, 2020), a website that collects and stores a large number of newspapers from around the world and classifies the articles by industry and sector. We have chosen the areas of Mining/Quarrying and Primary Metals Industries, to ensure that each news article concerns the sector under analysis.

Ideally, more than one newspaper per country would have been used in order to avoid or smooth out media biases. However, despite the fact that Factiva covers a broad range of newspapers, not all of them are available for the whole period under consideration and the number of newspapers covered is not the same for all the countries analysed. This data restriction meant that the period of time and the number of newspapers used to build the indices had to be limited. If more than one newspaper per country had been used, then the time period of available data would have been reduced to less than three years. We therefore decided to use only one newspaper per country and thus extend the time frame covered by the database to 2003–2016. As noted above, this is the database's major weakness, as the indices might be biased as a result of the different degrees of lobbying by the extractive industries, governments or other agents on the media with regard to coverage of conflicts in each country.

However, if the impact of lobbying remained the same over time, the indices in the country sample will only be biased on the country level. This is not a real concern because the variation will be still informative of increases or decreases in conflict levels. Problems will arise if the intensity of the lobbying varied over time. In the event that pressure from lobbyists varies across the country but remains stable over time, it will limit the country-level analysis, but it can be easily resolved by applying a fixed-effects regression model. However, steps should be taken in the future in an effort to avoid or minimize this potential pitfall.

The newspapers with the largest circulation were used for each country analysed, namely *El Mercurio* for Chile, *El Comercio* for Peru, the *Herald Sun* for Australia and *The Globe and Mail* for Canada. Data were available from November 2002 for *El Mercurio*, October 2002 for *El Comercio*, July 1997 for the *Herald Sun* and December 1986 for *The Globe and Mail*. Since most newspapers carry articles referring to other countries, we deleted those articles that contained the name of other countries and did not mention the name of the country of interest. The time period covering 2003 to the first semester of 2016 was chosen in order to have a balanced panel of data.

The identification of conflicts was crucial to the process. In line with UNDP/Fundación UNIR (2012), this study adopts a classic definition of social conflict as a process of contentious interaction between social actors and institutions which mobilize with different levels of organization and act collectively in order to improve conditions, defend existing situations, or advance new alternative social projects. A social conflict arises when a social group, actor or movement (workers, entrepreneurs, *campesinos*, indigenous peoples, teachers, civic movements, students, trade unions, academics, etc.) expresses a collective malaise in a hostile manner through demands and violent tactics designed to exert pressure (strikes, marches, riots, demonstrations, occupations of facilities, etc.) against any public or private body (p. 283).

Taking this definition, a dictionary-based approach was adopted to detect these patterns in the news articles. The hostile words category of the Harvard IV-4 dictionary[1] was used, with 687 entries in English at the time of writing. These were then translated into Spanish and the database was expanded to a total of 1,326 words by adding Spanish synonyms for those hostile words (see annex A1).

To avoid any possible complications arising from the conjugation of verbs and the agreement of nouns and adjectives, the common word endings in English and Spanish were removed using R software. Then, all words are rewritten in lower case letters, punctuation and accents were excluded, and common words in each language (like connectors) were deleted to avoid counting extra words that did not add any meaning and may be used more or less often in each language.

Lastly, the statistical program R allowed us to systematize this process to generate the indices for each country, which were then disaggregated by administrative area. Text mining techniques were subsequently carried out again; by checking to see if an article contained the name of the administrative area, the main cities and/or the mining sites in that area, the articles were categorized by region.

Three alternative indices were created that served as proxy variables for the number of social conflicts related to natural resources (Palazzo, n/d) and could be used for different purposes. They are:

- **Conflict news**: Let $CN_{i,t}$ be newspaper articles that include hostile words about the country or region i at time t, thus the conflict news index for each country or region is $CN_i=\sum_t CN_{i,t}$ and the total conflict news index for each period $CN_t=\sum_i CN_{i,t}$. This index may have numerous biases because, for example, it does not control for whether an increase in the number of hostile words is solely attributable to that fact that more reports are published on the matter.
- **Standardized conflict news**: In accordance with Baker, Bloom and Davis (2013), let $TN_{i,t}$ be newspaper articles about the extractive industries in country or region i published in period t, thus the standardized conflict news index is $SCN_{i,t}=CN_{i,t}/TN_{i,t}$.
- **Conflict intensity**: Based on García (2013) and Aromí (2013), the intensity of a conflict is measured at a specific point in time and space as the ratio of the number of hostile words to the total number of words inside the subset of conflict news. Let $CW_{i,t}$ be the number of hostile words found in the conflict news index and $TW_{i,t}$ the total number of words in those articles, thus conflict intensity in country or region i during the period t is $CI_{i,t} = CW_{i,t}/TW_{i,t}$.

III. Social conflicts related to extractive activities

1. Country comparisons

A total of 20,119 newspaper articles were collected about extractive industries from the four newspapers analysed covering the period between the first quarter of 2003 and the second quarter of 2016. Of these articles, 78% were classified as conflict news (see table 1),[2] indicating that the public is generally pessimistic about extractive activities. This was to be expected given that our unit of analysis is newspaper articles.

When the data are disaggregated by country, some differences arise. First, the percentage of standardized conflict news $(SCN_{i,t})$ is higher in the developed countries —around 92% for both Australia and Canada—, while in Chile and Peru it is between 63% and 66%. With regard to conflict intensity, the values are higher for the developed countries, with the figure for Canada slightly higher than that for Australia.

These results are descriptive of the general findings of this paper and the database that we have constructed. However, they should not be interpreted as meaning that conflict levels are higher

[1] See [online] http://www.wjh.harvard.edu/~inquirer/homecat.htm.

[2] Our main indices, disaggregated by month and year, are freely available at https://sites.google.com/view/gabrielmpalazzo/original-databases. Our codes in R format are available upon request.

in Australia and Canada than in Chile and Peru. Firstly, since we are comparing indices created from different newspapers, a higher proportion of standardized conflict news or conflict intensity could be explained by the preferences or interests of different readerships as well as the different writing styles of the newspapers' journalists. In addition, the differences in levels among the countries analysed coincide with languages differences; we therefore cannot discount the fact that native speakers of the two languages may express facts and their opinions in a different manner. Lastly, as was noted above, the differences might be attributable to different biases in the media. These results should therefore be interpreted as indices composed of different constants and the analysis should focus on how behaviour changes over time and the responses following relevant shocks from exogenous variables.

Table 1
Australia, Canada, Chile and Peru: national patterns in social conflicts related to extractive activities, first quarter of 2003–second quarter of 2016
(Total number of articles and percentages)

Country	Total number of articles on extractive activities $TN_{i,t}$	Conflict news $CN_{i,t}$	Standardized conflict news *(standardized index)* $SCN_{i,t}$ *(percentages)*	Conflict intensity $CI_{i,t}$ *(percentages)*
Australia	2 709	2 502	92.36	3.70
Canada	6 871	6 349	92.40	4.53
Chile	8 095	5 375	66.40	2.76
Peru	2 444	1 543	63.13	2.89
Total	20 119	15 769	78.37	3.81

Source: Prepared by the authors, on the basis of articles obtained from *El Mercurio*, *El Comercio*, *The Globe and Mail*, and *Herald Sun* newspapers.

Turning to time patterns (see table 2), three different subperiods can be detected with regard to the conflict news index, $CN_{i,t}$. The first, extending roughly from 2003 to 2006, where social conflict appears to be increasing; a second where conflict steadily decreases (2007–2009); and a third (2010–2016) where it remains generally stable (despite two spikes in 2010 and 2014). However, the standardized index $SCN_{i,t}$, reveals a different picture: the trend was relatively stable during the period 2003–2007, reflecting that conflict news grew *pari passu* with total news, before jumping and stabilizing at higher levels through the period 2008–2013; lastly, it fluctuates over the 2014–2016 period, peaking in 2015.

Table 2
Australia, Canada, Chile and Peru: time patterns in social conflicts related to extractive activities, 2003–2016
(Total number of articles and percentages)

Year t	Total number of articles on extractive activities $TN_{i,t}$	Conflict news $CN_{i,t}$	Standardized conflict news *(standardized index)* $SCN_{i,t}$ *(percentages)*	Conflict intensity $CI_{i,t}$ *(percentages)*
2003	822	695	84.55	4.02
2004	1 394	1 049	75.25	3.42
2005	1 497	1 091	72.88	3.20
2006	2 128	1 637	76.93	3.62
2007	1 952	1 270	65.06	3.63
2008	953	664	69.67	3.47
2009	774	529	68.35	4.02
2010	938	755	80.49	4.42
2011	670	495	73.88	4.03
2012	669	541	80.87	4.03
2013	732	535	73.09	3.98
2014	908	502	55.29	3.75
2015	908	697	76.76	3.91
2016	406	373	91.87	2.99

Source: Prepared by the authors, on the basis of articles obtained from *El Mercurio*, *El Comercio*, *The Globe and Mail*, and *Herald Sun* newspapers.

With regard to the conflict intensity, the $CI_{i,t}$ index indicates a steady increase in the intensity of social conflicts related to extractive activities between 2003 and 2010, before starting to diminish, albeit slowly.

Lastly, figure 1 shows the country-specific evolution of social conflict over time. Regarding both the total news index $TN_{i,t}$ and the conflict news index $CN_{i,t}$, it is clear that they peaked during the 2004–2008 period. In turn, the standardized index reveals that the aforementioned country differences are present over the whole period under analysis. The indices for Chile and Peru are more volatile than those of Australia and Canada, and social conflict in Chile grew significantly over the last two years of the period under analysis. Two interesting points arise with regard to conflict intensity. Firstly, that the Peru index is more volatile than the others; meaning that, unlike the other countries, Peru is subject to "explosions" of intense conflicts. Secondly, the conflict intensity remained relatively low in Chile during the whole period.[3]

Figure 1

Australia, Canada, Chile and Peru: social conflicts related to extractive activities, 2003–2016