The Positioning Matters

Estimating Geographical Bias in the Multilingual Record of Biographies on Wikipedia

Pablo Beytía*
Department of Social Sciences, Humboldt University of Berlin, Berlin, Germany
beytiapa@hu-berlin.de

ABSTRACT

This article proposes that an appropriate assessment of the geographical bias in multilingual Wikipedia's content should consider not only the *number* of articles linked to places, but also their *internal positioning* –i.e. their location in different languages and their centrality in the network of references between articles–. This idea is studied empirically, systematically evaluating the geographic concentration in the biographical coverage of globally recognized individuals (those whose biographies are found in more than 25 language versions of Wikipedia). Considering the internal positioning levels of these biographies, only 5 countries account for more than 62% of Wikipedia's biographical coverage. In turn, the inequality in coverage between countries reaches very high levels, estimated with a Gini coefficient of .84 and a Palma ratio of 207. In all the tests carried out, the inclusion of the linguistic and/or relational positioning of the articles increases the estimate of inequality in biographical coverage. This suggests that previous estimates of geographical bias, which do not consider differences in internal positioning, have underestimated the degree of inequality in the distribution of information.

CCS CONCEPTS

• **World Wide Web**; • **Information Storage systems**; • **Collaborative and social computing**;

KEYWORDS

Wikipedia, geographical bias, geo-tagged information, information inequality

ACM Reference Format:
Pablo Beytía. 2020. The Positioning Matters: Estimating Geographical Bias in the Multilingual Record of Biographies on Wikipedia. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3366424.3383569

1 INTRODUCTION

Along with the gender gap [1–4], geographical bias has been one of the most researched issues regarding inequalities in the content coverage of Wikipedia [5–10]. Although the distribution of georeferenced information is considerably different according to the language version of the encyclopedia [7, 8], studies show a transversal trend towards more content creation related to the United States and Western Europe, as well as a relative lack of information about some regions of Africa, the Middle East, Latin America and Asia [9, 10]. This trend is pretty widespread, and can be found both in the spatial distribution of the total number of geo-tagged articles [5], and in the distribution of subgroups of articles. For example, it can be seen in the storage of biographies of recognized persons –using their place of birth as a spatial approximation [11], in the coverage of relevant historical events –such as battles or wars–, and in the documentation of animals with established territorial origin –as is the case of reptiles [7]–. In short, today we know that Wikipedia tends to develop more information about people, animals, objects and events linked to the Global North and the traditional Western powers.

This geographical content gap, however, has typically been studied with three limitations. First of all, most articles analyze the inequality of coverage within specific languages, leaving aside the more global question of the *multilingual configuration* of encyclopedic information. Secondly, no *estimates of global information inequality* have been calculated so far, which would provide an approximation of the degree of geopolitical concentration in the recording and transmission of encyclopedic information. Finally, practically all studies use as an indicator of geographical coverage simply the number of articles associated with a territory, without considering that these articles do not have the same weight within this information system. The Wikipedia articles, in fact, have different degrees of *internal positioning* –due to their different levels of exposure in multiple languages or their degree of connectivity with other relevant articles– and this determines that they have different probabilities of dissemination and influence in the construction of discourses. As articles associated to places have specific levels of positioning, this phenomenon affects the degree of visibility or communicative centrality that those places have in encyclopedic information.

This article proposes that an appropriate evaluation of the geographical bias in the content of Wikipedia should not only consider the *number* of articles linked to places, but also the *internal positioning* of those articles –which could enhance or attenuate the visibility and centrality of information about places. To illustrate this idea, the geographical concentration in the biographies of globally recognized individuals –understood as those whose biography is available in more than 25 languages in Wikipedia [11] is analyzed. In addition to assessing the spatial inequality of this set of biographies (approximated by the birthplace of the recognized figures), here each article is weighted by its *Biographical Centrality*

*Research Fellow at the Alexander von Humboldt Institute for Internet and Society (HIIG), Guest Researcher at the Berlin Graduate School of Social Sciences (BGSS), and PhD Candidate at the Department of Social Sciences of Humboldt University of Berlin.

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383569

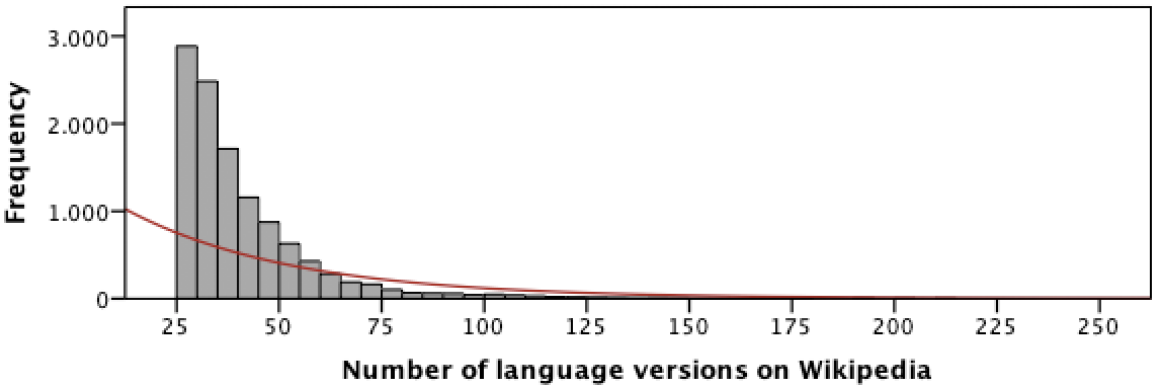

Figure 1: Distribution of the linguistic positioning of the biographies in more than 25 languages Source: Yu et al. 2016

Index (BCI) –an indicator of the degree to which each biography is exposed in different languages and its level of connectivity with other biographies [12, 13]. This allows approaching the differences of coverage in the content associated to the countries, without assuming (unjustifiably) that the articles are equivalent in terms of internal positioning.

2 METHODS

This The empirical objective of this study is to evaluate the geographic concentration of the multilingual biography record in Wikipedia by considering, simultaneously, the *number* of articles associated with the countries and the *internal positioning* of those articles. The methodology used considers the following aspects:

1. *Initial sample*: articles about "globally recognized" persons were analyzed, under the criterion that they have versions of their biography in more than 25 Wikipedia language versions. This information was obtained from Pantheon dataset 1.0 [11] and includes a total of 11,341 biographies. The geographical position associated with each article was approximated by the place of birth of the persons referenced.

2. *Linguistic positioning*: the biographies of the globally recognized figures could be in 26 *or more* languages out of the more than 300 available on Wikipedia. The number of languages in which each biography is located is a very relevant indicator of positioning, as it is associated with the amount of information, the inter-cultural coverage, the degree of dissemination and the discursive influence that each biography has. For instance, the biography of Marcelo Salas (former Chilean soccer player who has an article in 28 languages) is not equivalent to that of Confucius (Chinese philosopher and politician who has his biography in 192 languages). Figure 1 shows the distribution of the multilingual exposure of these biographies, and this information was extracted from Pantheon dataset 1.0 [11].

3. *Relational positioning*: Wikipedia articles are not configured as isolated websites, but as part of a system of articles that is established from references between websites (hyperlinks). These connections generate an information network where each article acquires a specific level of centrality. In Wikipedia in English, for example, the relational position of Klara Hitler (who receives no references from other relevant biographies) is not equivalent to that of Adolf Hitler (who receives hyperlinks from 340 biographies of well-known people). The "relational positioning" was calculated for each biography from its PageRank algorithm [14, 15] within the network of hyperlinks between biographies in English Wikipedia[1] (see Figure 2). This indicator –which can be understood as a measure of recursive centrality, since it gives greater weight to the biographical references that come from biographies with more biographical references– was obtained from the Networked Pantheon database [12].

4. *Biographical Centrality Index (BCI)*: for each biography, an indicator was calculated that includes its linguistic positioning (understood as the number of languages in which the article is available) and its relational positioning (or PageRank calculated in the network of references between biographies). BCI [12] "is an indicator of the degree of relevance of a biography in different languages, considering both its multilingual diffusion and its connectivity in the network of references generated within a specific group of biographies" [13]. Considering the number of languages of a biography (NL) and its PageRank (PR), the BCI of a famous character is summarized in the following formula:

$$\mathrm{BCI} = \frac{(\mathrm{NL} * \mathrm{PR}) - - (\mathrm{NL} * \mathrm{PR})\,\mathrm{min}}{(\mathrm{NL} * \mathrm{PR})\,\mathrm{max} - (\mathrm{NL} * \mathrm{PR})\,\mathrm{min}}$$

This indicator can be interpreted as a standardized measure of how likely it is that, within a group of biographies, a specific article is linked to the search for another article that is selected at random in a Wikipedia language also chosen randomly.

Although the BCI could be considered as an indicator of the importance of an article, it is based only on how Wikipedia organize the information (supply of content) and not on how much the articles are observed (demand for content). This significantly distinguishes the BCI from indicators based on the popularity of articles –such as Pageviews–, which in practice correlate very slightly with the BCI[2].

5. *BCI aggregation and inequality measures*: The concentration of biographical coverage in the countries was calculated by adding the BCI of the biographies of people born in their territory –or in other words: by multiplying in each country the number of biographies

[1]The hyperlink network was extracted from English Wikipedia, since that is the language version that covers the largest number of historical figures with biographies in more than 25 languages (11,340 out of a total of 11,341).
[2]Spearman's coefficient close to .4.

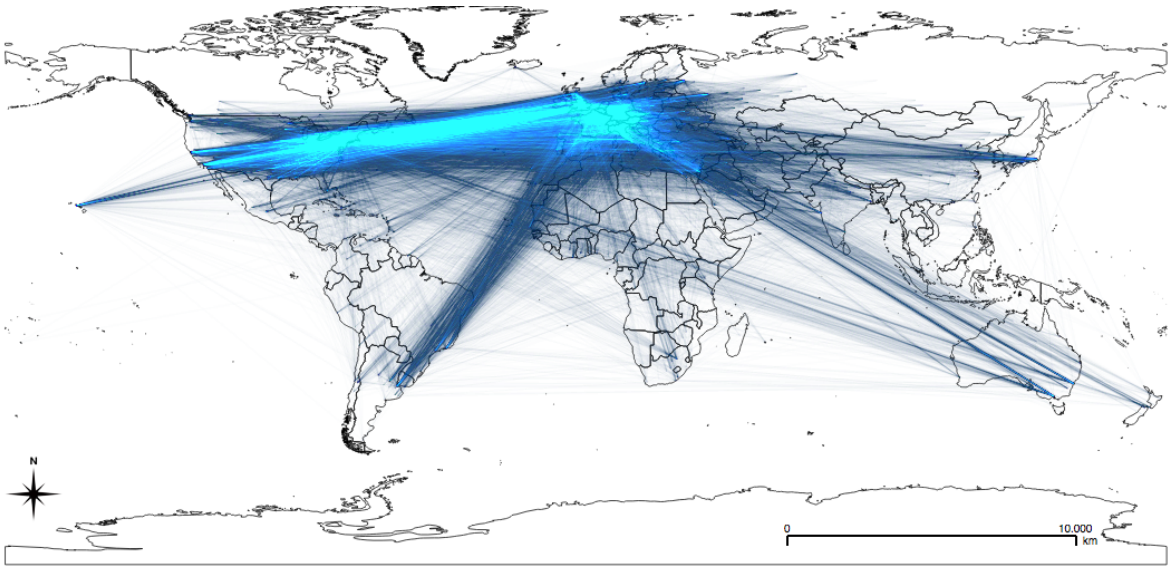

Figure 2: The network of hyperlinks between biographies in more than 25 languages

Note: references between biographies were obtained from English Wikipedia. Source: Beytía and Schobin 2018 [12]

Table 1: Top 10 countries with the greatest coverage of biographical information, ordered by accumulation of Biographical Centrality Index (BCI).

Country	Biographies	Biographies %	Linguistic positioning	Relational positioning	BCI	BCI %
United States	2169	19,13	87656	0,20694	22,263	22,67
United Kingdom	1147	10,11	47021	0,12734	12,986	13,22
Italy	808	7,12	36689	0,08785	9,541	9,72
France	867	7,64	35950	0,08391	8,352	8,51
Germany	748	6,60	30603	0,07164	7,860	8,00
Russia	374	3,30	15878	0,03329	3,993	4,07
Austria	140	1,23	6130	0,01790	2,529	2,58
Spain	296	2,61	12198	0,02333	2,305	2,35
Turkey	204	1,80	8695	0,01877	1,844	1,88
Poland	173	1,53	7320	0,01336	1,417	1,44

by the average BCI. Once the level of importance of each country in the biographical record was established, the usual indicators of inequality were calculated, among which the Gini coefficient [16] and the Palma ratio stand out [17, 18].

3 RESULTS

From a broad perspective, the geographical structure of inequality in information coverage is consistent with previous studies: the distribution of BCI is concentrated in the Global North, and especially in the United States, the United Kingdom, Italy, France and Germany (see Figure 3 and Table1). The degree of concentration of the biographical coverage is such that these 5 countries contain 50.6% of the total biographies available in more than 25 languages and 62.1% of the biographical coverage if the positioning of the articles is considered (see BCI % in Table 1). In other words, the content concentration is very high when observing the spatial distribution of the biographies, but it is even higher when considering the linguistic and relational positioning of the articles.

A similar situation can be observed when estimating global inequality indicators: geographical inequality is higher when considering the internal positioning of items. By simply comparing the number of biographies of people born in the countries –as previous studies have done–, a Gini coefficient of .79 and a Palma ratio of 41 are established (i.e.: 10% of the countries with the highest coverage have 41 times the number of biographies recorded by 40% of the countries with the lowest coverage). Both measures indicate a high level of information inequality at the global level. However, when considering also the positioning of the articles (see BN * BCI, in Table 2) this inequality is even greater, obtaining a Gini coefficient of .84 and a Palma ratio of 207 (which is 5 times higher than that obtained without considering the positioning).

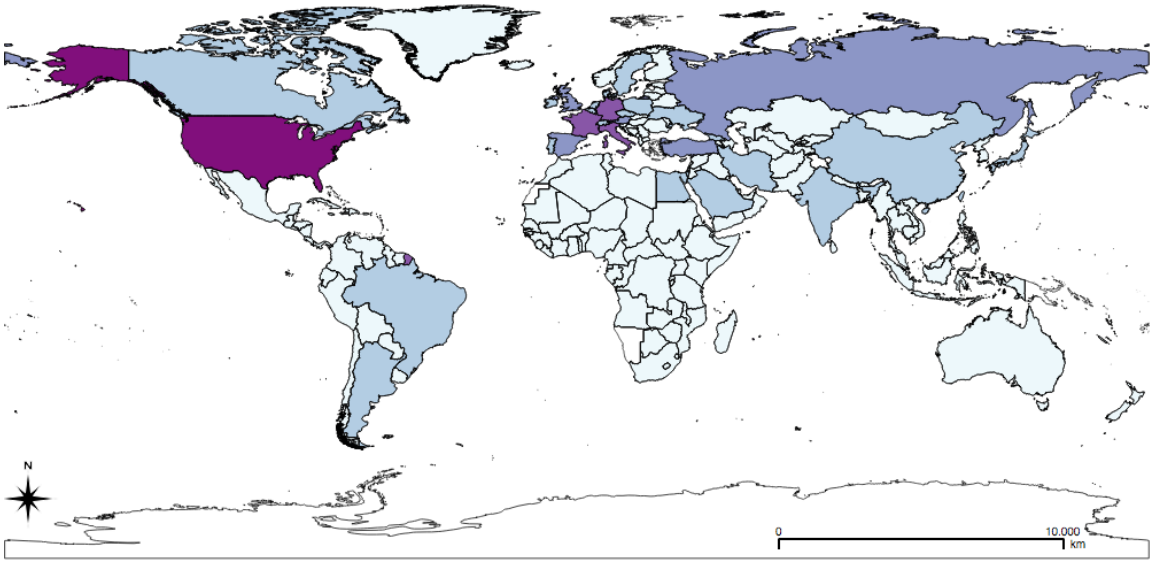

Figure 3: Spatial concentration of the biographical coverage in Wikipedia considering the internal positioning of the articles (BCI accumulation)

Note: the darker the color, the more biographical information coverage the country has (number of people who have their biography in more than 25 languages, multiplied by the average Biographical Centrality Index in the country).

Table 2: Measures of inequality in coverage associated with countries

	Gini	10/40 (Palma ratio)	10/10
BN (biography number)	.79	41	232
BN * language positioning	.80	56	564
BN * relational positioning	.83	118	2601
BN * BCI	.84	207	7677

4 CONCLUSION

This article explored the geographical bias in Wikipedia's coverage of the biographies of globally recognized individuals. Unlike previous studies on geographical bias, territorial coverage was evaluated not only by the number of articles distributed in the territory, but also by the "internal positioning" of those articles within this information system. To approach this internal positioning, the Biographical Centrality Index (BCI) was employed –an indicator of the relevance of each biography, which considers the number of languages in which the articles are available and their level of centrality in the reference network between articles–. The accumulation of BCI in the countries was used to measure their degree of biographical coverage, and based on that, general indicators of concentration and inequality of biographical content in Wikipedia were generated.

While the geographical distribution of content points in the same direction as previous studies –that is, a concentration of information about people who were born in the Global North or in the traditional Western powers– this article calculates more precise indicators of concentration and inequality at the global level (Gini coefficient, Palma ratio and 10/10 ratio), which allow the geopolitical inequality of information to be systematically assessed. The main results can be classified into three observations:

1. *On concentration:* biographical coverage is highly concentrated in 5 countries –the United States, the United Kingdom, Italy, France and Germany– where more than half of the people with biographies available in more than 25 languages were born. These countries also account for 62.1% of biographical coverage when considering the internal positioning (BCI) of biographies.
2. *On inequality*: the global Gini coefficient of biographical coverage was estimated between .79 and .84, depending on whether the positioning of the biographies in the calculation is considered and what type of positioning is used as a reference. Similarly, the Palma ratio varies between 41 and 207, according to the method adopted to distinguish the positioning of the articles (see table 2). Although data comparability needs to be reviewed in more detail, this level of inequality appears to be higher than the global distribution of wealth [19, 20], and similar to the distribution of land in the most unequal regions of the world (such as Latin America) [21].

3. *On the effect of positioning in the evaluation of geographical information coverage:* regardless of the indicator chosen (linguistic, relational or BCI), consideration of the positioning of the articles increased the estimate of information concentration and inequality in coverage in all the tests. This means that global inequality of biographical content has probably been underestimated, since previous studies have assumed that internal positioning was not relevant for estimating differences in coverage between territories. As shown here, consideration of internal positioning not only generates significant differences in the estimation of spatial inequality of content, but also tends to show that the information is (even) more concentrated in a few Western countries of the Global North.

In addition to these results, this research has proposed a specific methodology to consider the internal positioning of articles in the evaluation of Wikipedia's geographical bias. This methodology, however, is not only applicable to the evaluation of *geographical* inequalities, but also to the estimation of information bias of any kind. It would be interesting for future research to replicate this methodology for the evaluation of other relevant information inequalities in Wikipedia –as is the case with the gender gap– and to assess whether in those cases the accumulation of internal positioning also establishes an amplification of the levels of concentration and inequality of information.

Finally, future research could also contribute to completing this methodology with new tests and variables. This article can be considered as a first step, which also opens a rather complex discussion. The main argument has been that the inequalities of information coverage are not only related to the *number* of articles linked to territories or social groups, but also to the *position* that these articles have within a complex information system. However, what has been called "internal positioning" here is potentially only one of the factors that could be participating in the construction of these biases in Wikipedia. For example, differences in the *quality* of the articles –considering variables such as the amount of text, the use of images or the selection of sources– could also amplify the differences in coverage over territories or social groups. In that sense, future studies could explore the inclusion of new factors associated with inequality of coverage, and advance in the generation of a more complete model to estimate information biases in Wikipedia.

ACKNOWLEDGMENTS

The author is grateful for a doctoral scholarship granted by CONICYT (National Commission for Scientific and Technological Research of the Chilean Government) and the German Academic Exchange Service (DAAD).

REFERENCES

[1] Gruwell, L. Wikipedia's politics of exclusion: Gender, epistemology, and feminist rhetorical (in) action. *Computers and Composition* **37**, 117–131 (2015).
[2] Klein, M., Gupta, H., Rai, V., Konieczny, P. & Zhu, H. Monitoring the Gender Gap with Wikidata Human Gender Indicators. in *Proceedings of the 12th International Symposium on Open Collaboration* 1–9 (2016).
[3] 3Shane-Simpson, C. & Gillespie-Lynch, K. Examining potential mechanisms underlying the Wikipedia gender gap through a collaborative editing task. *Computers in Human Behavior* **66**, 312–328 (2017).
[4] Hinnosaar, M. Gender inequality in new media: Evidence from Wikipedia. *Journal of Economic Behavior & Organization* **163**, 262–276 (2019).
[5] Graham, M., Hogan, B., Straumann, R. K. & Medhat, A. Uneven geographies of user-generated information: patterns of increasing informational poverty. *Annals of the Association of American Geographers* **104**, 746–764 (2014).
[6] Graham, M. Information geographies and geographies of information. *New geographies* (2015).
[7] Roll, U. *et al.*Using Wikipedia page views to explore the cultural importance of global reptiles. *Biological conservation* **204**, 42–50 (2016).
[8] Overell, S. E. & Rüger, S. View of the world according to Wikipedia: Are we all little Steinbergs? *Journal of Computational Science* **2**, 193–197 (2011).
[9] Graham, M., Hale, S. A. & Stephens, M. Geographies of the World's Knowledge. (2011).
[10] Graham, M., De Sabbata, S. & Zook, M. A. Towards a study of information geographies:(im) mutable augmentations and a mapping of the geographies of information. *Geo: Geography and environment* **2**, 88–105 (2015).
[11] Yu, A. Z., Ronen, S., Hu, K., Lu, T. & Hidalgo, C. A. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data* **3**, 150075 (2016).
[12] Beytía, P. & Schobin, J. Networked Pantheon: a Relational Database of Globally Famous People. *Available at SSRN 3255401* (2018).
[13] Beytía, P. & Müller, H.-P. Towards a Digital Reflexive Sociology: Exploring the Most Globally Disseminated Sociologists on Multilingual Wikipedia. (2019).
[14] Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* **30**, 107–117 (1998).
[15] Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank citation ranking: Bringing order to the web.* (1999).
[16] Gini, C. Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1912).
[17] Palma, J. G. Homogeneous middles vs. heterogeneous tails, and the end of the 'inverted-U': It's all about the share of the rich. *development and Change* **42**, 87–153 (2011).
[18] Palma, J. G. Do nations just get the inequality they deserve? The "Palma Ratio" re-examined. in *Inequality and Growth: Patterns and Policy* 35–97 (Springer, 2016).
[19] Hellebrandt, T. & Mauro, P. The future of worldwide income distribution. *Peterson Institute for International Economics Working paper* (2015).
[20] Darvas, Z. *Some are more equal than others: new estimates of global and regional inequality.* (IEHAS Discussion Papers, 2016).
[21] Guereña, A. Unearthed: land, power, and inequality in Latin America. *Oxfam International* (2016).