

Preliminary results of Wikipedia readership research: Regional self-focus and relatively stable viewing of broad topic areas across regions

Andrew Russell Green
Independent/unaffiliated
andyrussg@gmail.com

Introduction

This paper describes the preliminary results of an investigation of patterns in Wikipedia pageviews across geographic regions. Our primary finding to date is that readers demonstrate substantial self-focus in the geographic regions associated with the articles they view. A secondary finding is that the proportions of broad topic areas (e.g., media, science, geography) associated with pageviews vary less across regions compared to the variations observed in geographic associations. These results are based on samples of pageviews from October 2025. Below, we detail these findings and describe the methods we used to obtain them. Work to complete this analysis is ongoing; we also summarize next steps.

This research is part of a larger exploratory project to study Wikimedia communities, readership and content using causal and social systems approaches. In addition to advancing understanding of those domains, we seek to develop theory and methods for linking micro-level human activity observed in trace data to meso- and macro-level social phenomena (MetaWiki, 2026).

The research questions for the investigation whose results we describe here (the current stage of the larger project) are as follows:

RQ1: What patterns exist in Wikipedia pageviews, related to the topics and locations *associated with* the articles viewed, across *readers' own* geographic locations and Wikipedia language editions?

(Here, “topics and locations *associated with* the articles” means, “topics and locations that the articles are about, or that are closely related to the articles.” For example, an article about a scientist born in country A would be associated with both science and country A.)

RQ2: What hypotheses about social phenomena could explain these patterns?

RQ3: What models can we build to investigate these hypotheses?

The results we present are primarily related to RQ1. To obtain them, we analyzed data from the Wikimedia Foundation’s public Differential Privacy Pageviews (DPP) dataset (Wikimedia Foundation, 2026) in conjunction with the output of machine learning models that predict articles’ geographic and topic associations, and auxiliary data on global internet access rates.

While imbalances in available Wikipedia content have received considerable attention (e.g., Oxford Internet Institute, 2018), patterns of topic and geographic associations in content consumption have been comparatively understudied.

Data preparation and analysis

The DPP dataset is the most detailed public source for Wikipedia pageview data, containing daily per-country pageview counts for nearly all articles across all Wikipedia language editions. The data is aggregated from the logs of Wikimedia Foundation (WMF) servers. Device country is determined using IP-based geolocation.

To protect reader privacy, the WMF adds noise and removes information about some articles and countries. Even so, the dataset offers a usable reflection of reader activity for most countries.

Due to the size of the dataset, it was not possible to fetch the machine learning models’ predictions for all articles viewed over a reasonable timespan, so pageviews were sampled, and models were queried for the associations of articles appearing in the samples.

The rows of the DPP tables are articles, not pageviews. However, the pageview is our main unit of analysis. To sample pageviews by country, we virtually de-aggregate the DPP tables as follows: we calculate the total pageviews (N) in the dataset for each country, including pageviews from all Wikipedia language editions, over the selected timespan (October 2025). We then generate n random integers between 0 and $N-1$, and retrieve information about each sampled pageview for the country from the DPP files as if the files contained rows of country-specific pageviews indexed from 0 to $N-1$, taking the random numbers to be indices of those virtual rows. (This is formally equivalent to sampling without replacement from a table where rows are pageviews.)

For the topic and geographic associations of the articles viewed, we queried models about English-language versions of articles whenever possible. We used the Wikidata QIDs and the Wikidata API to find equivalent page titles in English Wikipedia.

For articles with no QID, we queried the machine learning models about the article in the language it was originally viewed in.

The Wikipedia API was also consulted to determine if articles were disambiguation pages or if they had been deleted since being viewed.

The code for the data pipeline, calculations and visualizations was written in Python and is available online.

The model we used for articles’ geographic associations outputs predictions for country associations. These were processed into binary region associations based on the region that each country was considered to be a part of according to a modified United Nations Level 2 region scheme. (This scheme was selected as a temporary heuristic for exploratory purposes; although the country groupings it defines are widely used, it is not a data-based alignment of countries with regional social subsystems.)

We also aggregated the granular output of the model for topic associations to obtain five non-exclusive top-level topic categories: Culture, Geography, History and society, Media and STEM. This follows the approach used by previous research on temporal patterns in pageviews of article topic areas (Piccardi et al., 2024).

We attempted to sample 2500 pageviews per country, however, for some countries, this exceeded the number of pageviews in the dataset. We set the following cutoffs for sample validity: a minimum effective sample size of 200 (defined as the number of pageviews in the sample that were of articles with valid model outputs) and a minimum of 40 unique articles in the effective sample.

Thus, for each country with valid data (i.e., countries included in the DPP dataset and for which we had valid samples), we obtained estimates of the proportions of pageviews associated with each region and topic area. To calculate a global baseline and regional averages, we weighted the estimates for each country by the country’s total internet-connected population. This partly offset the bias from countries that have relatively more pageviews but smaller populations. All the estimated pageview association proportions shown here use this weighting and omit countries without valid data. Internet-connected population data was downloaded from the World Bank (World Bank Group, 2026).

We also calculated divergence scores for each estimated proportion, defined as $\log_2(r_p/b_p)$ (Log Risk Ratio), where r and b are the weighted-average estimates for proportion p for the region and global baseline, respectively. Margins of error (not shown in figures) were calculated at 95% confidence.

Results and discussion

Figures 2 and 3 show internet-population-weighted proportion estimates and divergence scores for pageviews’ regional association, by region in which the device requesting the pageview was deemed to be located. Figures 4 and 5 show the same information for pageviews’ topic

area associations. (An overview of data validity and population is omitted due to lack of space. For nearly all regions, we have valid data covering nearly all of the population.)

We speculate that causes of regional self-focus might include reader preference for content related to a local identity, or search engine results favoring local content.

The lower variation in the proportions of broad topic areas could imply that the contexts and motivations for accessing Wikipedia are globally relatively invariant; this result also connects to previous research (Lemmerich et al., 2019).

Next steps

Planned next steps related to the results described above include modeling noise and missing data due to differential privacy processing, refining the sampling approach, and extending the analysis to a longer time period. We will also conduct sensitivity analyses to validate the cutoffs for minimum effective sample size and unique articles, and calculate e-values to assess the robustness of our conclusions.

To begin addressing RQ2 and RQ3, we will conduct a granular topic analysis of the top articles contributing to geographical self-focus and develop models for clustering countries based on patterns in reader focus. We hope these models will allow us to replace the UN Level 2-based region scheme and distinguish country-level self-focus from regional self-focus signals.

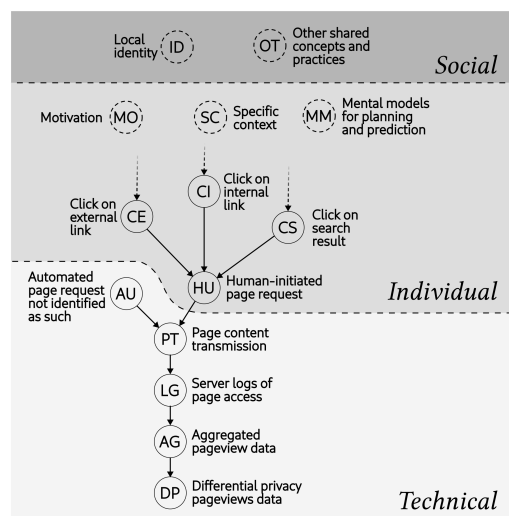


Figure 1: Diagram showing causes and effects considered in this analysis. The technical layer includes well-defined causes and effects, and the individual (reader) layer has some well-defined causes and effects. However, causes and effects that emerge from complex interactions at the social layer are typically difficult to define, though they can be considered in qualitative terms, and we can form hypotheses about their appearance in statistical models.

Pageview device location	Est. pageview association proportions											
	Central Asia	Eastern Asia	Eastern Europe	Lat. Am. and the Caribbean	Northern Africa	Northern America	Oceania	South-Eastern Asia	Southern Asia	Sub-Saharan Africa	Western Asia	Western Europe
Central Asia	16.65%	2.03%	4.42%	2.17%	0.05%	4.31%	0.70%	0.06%	2.84%	0.07%	2.86%	5.30%
Eastern Asia	0.04%	73.88%	0.60%	0.93%	0.17%	8.05%	0.35%	0.93%	0.32%	0.41%	0.35%	4.11%
Eastern Europe	0.71%	1.45%	41.27%	3.11%	0.43%	13.79%	0.69%	0.71%	0.83%	0.79%	2.70%	12.07%
Lat. Am. and the Caribbean	0.14%	2.34%	0.83%	42.70%	0.51%	20.80%	0.76%	0.66%	0.73%	0.98%	1.40%	10.43%
Northern Africa	0.02%	0.47%	0.37%	2.04%	29.73%	3.67%	0.36%	0.22%	0.21%	0.69%	5.67%	9.27%
Northern America	0.04%	3.43%	1.45%	3.49%	0.54%	62.21%	1.87%	0.97%	1.39%	1.19%	1.45%	13.76%
Oceania	0.19%	2.84%	1.44%	2.35%	0.22%	40.30%	19.52%	1.07%	1.97%	1.51%	2.06%	20.83%
South-Eastern Asia	0.21%	8.94%	1.05%	2.52%	0.62%	12.48%	1.53%	41.92%	1.45%	0.99%	2.04%	9.73%
Southern Asia	0.07%	2.34%	0.93%	1.24%	0.38%	10.02%	1.28%	1.38%	50.30%	1.13%	1.43%	6.32%
Sub-Saharan Africa	0.02%	0.79%	0.48%	2.43%	0.47%	10.02%	0.68%	0.55%	0.67%	30.05%	0.56%	9.19%
Western Asia	0.03%	0.57%	0.76%	1.98%	1.97%	6.68%	0.36%	0.53%	4.97%	0.87%	22.10%	4.67%
Western Europe	0.21%	2.13%	2.20%	3.58%	1.05%	23.32%	1.48%	0.64%	1.04%	1.54%	2.26%	53.90%

Figure 2: Estimated weighted proportions of regional associations of pageviews, by device region.

Pageview device location	Divergence score for region											
	Central Asia	Eastern Asia	Eastern Europe	Lat. Am. and the Caribbean	Northern Africa	Northern America	Oceania	South-Eastern Asia	Southern Asia	Sub-Saharan Africa	Western Asia	Western Europe
Central Asia	5.41	-1.61	0.25	-1.82	-5.15	-2.06	-0.84	-6.45	-2.29	-5.62	0.23	-1.37
Eastern Asia	-3.32	3.57	-2.64	-3.04	-3.43	-1.16	-1.84	-2.54	-5.44	-3.07	-2.81	-1.74
Eastern Europe	0.86	-2.10	3.47	-1.30	-2.12	-0.38	-0.86	-2.93	-4.06	-2.13	0.14	-0.18
Lat. Am. and the Caribbean	-1.54	-1.41	-2.17	2.48	-1.88	0.21	-0.72	-3.02	-4.26	-1.81	-0.80	-0.39
Northern Africa	-4.50	-3.72	-3.32	-1.91	3.99	-2.29	-1.80	-4.59	-6.04	-2.33	1.21	-0.57
Northern America	-3.15	-0.86	-1.36	-1.13	-1.81	1.79	0.58	-2.48	-3.32	-1.53	-0.76	0.01
Oceania	-1.07	-1.13	-1.36	-1.71	-3.10	1.17	3.96	-2.34	-2.82	-1.20	-0.25	0.60
South-Eastern Asia	-0.88	0.52	-1.83	-1.60	-1.59	-0.53	0.30	2.96	-3.26	-1.80	-0.26	-0.50
Southern Asia	-2.56	-1.41	-1.99	-2.62	-2.30	-0.84	0.03	-1.97	1.86	-1.61	-0.77	-1.12
Sub-Saharan Africa	-4.64	-2.98	-2.97	-1.66	-1.98	-0.84	-0.87	-3.30	-4.36	3.12	-2.12	-0.58
Western Asia	-3.53	-3.44	-2.29	-1.95	0.07	-1.43	-1.78	-3.34	-1.48	-1.98	3.18	-1.56
Western Europe	-0.91	-1.54	-0.76	-1.10	-0.84	0.38	0.24	-3.08	-3.73	-1.16	-0.11	1.97

Figure 3: Divergence scores for proportions of regional associations of pageviews, by device region.

	Culture	Geography	History and society	Media	STEM
Central Asia	50.25%	60.67%	15.15%	10.48%	22.17%
Eastern Asia	60.94%	71.42%	10.44%	42.56%	8.05%
Eastern Europe	50.35%	65.25%	21.67%	39.27%	16.81%
Lat. Am. and the Caribbean	62.26%	57.58%	11.58%	29.95%	9.97%
Northern Africa	61.48%	58.58%	21.18%	25.37%	19.93%
Northern America	55.00%	44.47%	14.81%	45.86%	9.45%
Oceania	58.44%	53.59%	12.09%	48.80%	7.69%
South-Eastern Asia	53.82%	68.07%	12.93%	32.14%	12.64%
Southern Asia	44.81%	69.23%	9.84%	54.97%	12.93%
Sub-Saharan Africa	51.08%	56.65%	14.51%	27.98%	20.29%
Western Asia	57.04%	59.99%	16.53%	30.34%	23.78%
Western Europe	57.43%	57.28%	14.29%	39.05%	8.19%

Figure 4: Estimated weighted proportions of topic area associations of pageviews, by device region.

	Culture	Geography	History and society	Media	STEM
Central Asia	-0.09	-0.02	0.18	-1.93	0.76
Eastern Asia	0.19	0.21	-0.36	0.10	-0.70
Eastern Europe	-0.09	0.08	0.70	-0.02	0.36
Lat. Am. and the Caribbean	0.22	-0.10	-0.21	-0.41	-0.39
Northern Africa	0.20	-0.07	0.66	-0.65	0.61
Northern America	0.04	-0.47	0.15	0.20	-0.47
Oceania	0.13	-0.20	-0.14	0.29	-0.77
South-Eastern Asia	0.01	0.14	-0.05	-0.31	-0.05
Southern Asia	-0.25	0.17	-0.44	0.46	-0.02
Sub-Saharan Africa	-0.06	-0.12	0.12	-0.51	0.63
Western Asia	0.09	-0.04	0.31	-0.39	0.86
Western Europe	0.10	-0.11	0.10	-0.03	-0.68

Figure 5: Divergence scores for proportions of topic area associations of pageviews, by device region.

References

[Lemmerich et al.2019] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the world reads Wikipedia: Beyond English speakers. *WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*.

[MetaWiki2026] MetaWiki. 2026. Research: Exploring Wikimedia communities, trace data, social systems and causality. https://meta.wikimedia.org/wiki/Research:Exploring_Wikimedia_Communities,_Trace_Data,_Social_Systems_and_Causality.

[Oxford Internet Institute2018] Oxford Internet Institute. 2018. The uneven geography of Wikipedia. <https://geography.oii.ox.ac.uk/the-uneven-geography-of-wikipedia/>.

[Piccardi et al.2024] Tiziano Piccardi, Martin Gerlach, and Robert West. 2024. Curious rhythms: Temporal regularities of Wikipedia consumption. *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media. Association for the Advancement of Artificial Intelligence*.

[Wikimedia Foundation2026] Wikimedia Foundation. 2026. Pageviews differential privacy — current. https://analytics.wikimedia.org/published/datasets/country_project_page/00_README.html.

[World Bank Group2026] World Bank Group. 2026. Individuals using the internet (% of population). <https://data.worldbank.org/indicator/IT.NET.USER.ZS>.