Which Factors are Associated with Wikipedia Relevancy?*

An Analysis of Notable People on Wikipedia

Parth Samant

November 27, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Understanding prominence of individuals on offers a way to quantify the societal impact of individuals through time. By using a dataset of notable people on Wikipedia from 3500 BCE to 2018 AD, this paper uses a bayesian linear model to identify certain factors associated with prominence. Some of the factors focused on include nationality, gender, occupation, and the number of years since the individuals birth. Analysing potential trends allows us to evaluate potential patterns and bias in one's relevance.

Estimand paragraph

The estimating (or what we are estimating) is the prominence of their Wikipedia biography based on percentile. "Prominence" of a biography is determined using multiple metrics, such as the average amount of views per year, the total word count, and the number of Wikipedia editions.

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

*Code and data are available at: https://github.com/samantparth/Wikipedia-Historical-Prominence-Trends.

2 Data

2.1 Overview

All data analysis was done through the statistical programming language R (R Core Team 2023) with the packages `tidyverse`(Wickham et al. 2019), `rstanarm` (Goodrich et al. 2022), `modelsummary` (Arel-Bundock 2022), `arrow`(Richardson et al. 2024), `readr`(Wickham, Hester, and Bryan 2024), `httr`(Wickham 2023), `R.utils`(Bengtsson 2023), `knitr`(Xie 2024) and `ggplot2` (Wickham 2016).

2.2 Measurement ?@sec-measurement

Wikipedia is an online encyclopedia that provides information on numerous subjects, including the lives of those who are relatively well-known.Thus, the data comes from a research study (Laouenan et al. 2022) that aims to build a cross-verified database of 'notable people' who have ever lived by using information from Wikipedia. Those with a Wikipedia article are considered "notable", as the overwhelming majority of people who have ever lived do not have one. Additionally, this dataset of notable people is very large (with hundreds of thousands of entries), reflecting the large amount of notable people that are written about on Wikipedia.

This data was obtained using the Wikidata universe (which provides data on Wikipedia), where they used the "instance of humans" category to select for a sample of notable individuals. Many potential obstacles (including identifying non-humans vs humans or multiple biographies) were controlled for when this dataset was made.

Based on the information provided in those Wikipedia articles, the study was also able to identify characteristics of these people. Some of these factors include nationality, occupation, and age.

Furthermore, the researchers constructed many variables including `ranking_visib_5criteria`, which ranks the prominence of an individual on Wikipedia. The ranking is dependent on these 5 metrics :

- the number of different editions;
- the number of non-missing items for birth date, gender, and domain of influence;
- the total number of words for their article;
- average yearly number of viewers from 2015 to 2018;
- number of external links (such as sources and references) from Wikidata.

2.3 Data Cleaning and Variables

This dataset was cleaned by first mutating some variables, selecting/renaming variables of interest, and filtering out rows with missing information, randomly selecting 10,000 entries from the dataset. The variables selected are mentioned later on in this section.

Since this paper focuses on predicting historical prominence, I chose variables which I thought could be associated with the outcome variable (`percentile_rank`).

Outcome Variable:

- `**percentile_rank**`: a transformation of the variable `ranking_visib_5criteria` (as mentioned in **?@sec-measurement**). This indicates the percentile associated with the notability ranking. A higher percentile indicates they were a more notable person.

Predictor Variables:

- `**subregion**`: the UN subregion corresponding to where they were from. This is simply a renaming of the variable `un_subregion`.
- `**years_since_birth**`: the number number of years that have passed since their birth. If they are alive, this is simple their age.
- `**time_period**`: The time period that they were born in.
- `**gender**`: The reported gender of the individual (either male or female).
- `**occupation**`: The primary field/occupation that the individual is known for.

2.4 Analysis of Variables

2.4.1 Subregion

In **?@fig-subregion-counts**, we can visualize the distribution of each geographic sub region.

```
#| label: fig-subregion-counts
#| fig-cap: Proportion of Notable People by Subregion
#| echo: false


summary_table_subregion <- cleaned_data |>
  group_by(subregion) |>
  summarize(Count = n()/nrow(cleaned_data)) |>
  arrange(desc(Count))
```

```
summary_table_subregion |>
 kable()
```

subregion	Count
Western Europe	0.4584
Northern America	0.1554
Southern Europe	0.0873
South America	0.0584
Northern Europe	0.0522
Eastern Europe	0.0469
Eastern Asia	0.0369
Oceania Western World	0.0248
South Asia incl. Indian Peninsula	0.0194
Western Asia (Middle East Caucasus)	0.0152
Central America	0.0100
SouthEast Asia	0.0094
West Africa	0.0052
Southern Africa	0.0048
Caribbean	0.0045
North Africa	0.0045
East Africa	0.0032
Central Asia	0.0015
Central Africa	0.0013
Oceania not Aus Nze	0.0007

?@fig-subregion-counts provides a detailed view of the prevalence of notable people by subregion. Interestingly, subregions that are a part of the Western world tend to be over-represented in terms of notable people. In fact, the difference is so stark that the most well-represented region (Western Europe) has nearly 100x the representation of notable people than West Africa.

There could be many explanations for this, such as how Western nations tend to have more global cultural dominance (and thus more 'notable people').

2.4.2 time_period/years_since_birth

These are two variables that are directly associated with each other, since the time period of ones life is directly a result of the number of years since their birth.

However, a notable feature of `years_since_birth` (and thus `time_period`) is the distribution as shown in Figure 1.

Figure 1: Distribution of Notable Individuals by Years Since Birth

From Figure 1, we can see that the overwhelming majority of documented notable people on Wikipedia tend to be born in the past 500 or so years. This makes sense, as retrieving information on individuals that lived a longer time ago is more difficult. Future plots and analyses will thus use the logarithm of this variable to have a more detailed view on its impact.

2.4.3 Time Period and Subregion

Different regions had different levels of cultural relevance/dominance depending on the time period. Because of this fact, it may be possible that the prevalence of subregions can heavily depend on the number of years since their birth.

In Figure 2, we can see this relationship quite clearly. Note that the proportions are out of the top 5 most popular subregions, rather than every subregion. This was done to ensure readability.

From Figure 2, we can confirm the relationship between the time period and subregion - that the prevalency of notable individuals heavily changes depending on the time period in question. The most notable example is the decrease in prevalence for Southern Europe/East Asia and the subsequent rise of Northern America (i.e., the USA and Canada) and Western Europe.

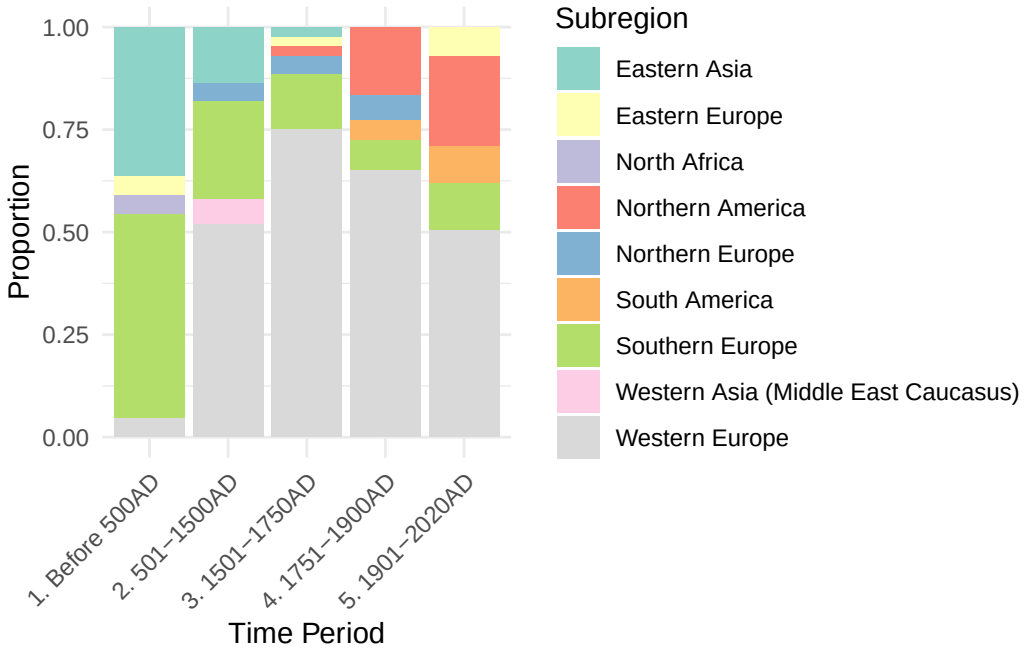

Figure 2: Top 5 Subregions by Time Period

As mentioned above, this confirms the phenomenon of how different subregions have different levels of influence depending on the time period. This graph especially highlights the rise of notable people in Western Europe as well as its descendant countries.

2.4.4 Gender and Occupation

The remaining variables, gender and occupation, may also show some sort of trend (since ones field of work is often quite gendered). This relationship is shown in **?@tbl-gender-occupation**, which shows ones occupation (or more specifically, what they are known for).

occupation	proportion_female	proportion_male
Nobility	0.34	0.66
Family	0.29	0.71
Other	0.26	0.74
Culture-core	0.24	0.76
Culture-periphery	0.21	0.79
Missing	0.18	0.82
Worker/Business (small)	0.14	0.86
Sports/Games	0.12	0.88
Politics	0.11	0.89

occupation	proportion_female	proportion_male
Academia	0.11	0.89
Corporate/Executive/Business (large)	0.08	0.92
Administration/Law	0.06	0.94
Explorer/Inventor/Developer	0.04	0.96
Religious	0.03	0.97
Military	0.02	0.98

Occupations of Notable People by Gender

?@tbl-gender-occupation shows that those of nobility (often by being born in a high social rank) tend to have a higher proportion of females, where other occupations - such as the military - tend to have a higher proportion of males.

Interestingly, even the most female-dominated occupation (nobility) is still roughly 2/3rds men. This reflects a clear bias in people that are considered 'notable': men are very much over-represented in every occupation (and in this dataset, as a result).

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

Table 3: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	1.12
	(1.70)
length	0.01
	(0.01)
width	−0.01
	(0.02)
Num.Obs.	19
R2	0.320
R2 Adj.	0.019
Log.Lik.	−18.128
ELPD	−21.6
ELPD s.e.	2.1
LOOIC	43.2
LOOIC s.e.	4.3
WAIC	42.7
RMSE	0.60

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 3.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

References

Arel-Bundock. 2022. "Modelsummary: Data and Model Summaries in r." *Journal of Statistical Software*. https://doi.org/10.18637/jss.v103.i01.

Bengtsson, Henrik. 2023. *R.utils: Various Programming Utilities*. https://CRAN.R-project.org/package=R.utils.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Laouenan, Morgane, Palaash Bhargava, Jean-Benoît Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. 2022. "A Cross-Verified Database of Notable People, 3500BC-2018AD." *Scientific Data* 9 (1): 290. https://doi.org/10.1038/s41597-022-01369-4.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Httr: Tools for Working with URLs and HTTP*. https://CRAN.R-project.org/package=httr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.