

Bayesian (comparative) journalism studies

Chung-hong Chan¹ & Adrian Rauschfleisch²

¹ Mannheimer Zentrum für Europäische Sozialforschung, Universität Mannheim, Germany

² National Taiwan University, Taiwan

Author Note

The authors would like to thank professor Kasper Welbers (Vrije Universiteit Amsterdam) for his advice on the development of the Dutch dictionary; Dr Junior Yuner Zhu (City University of Hong Kong) for sharing details about her study. Source code and data are available at (redacted). Prereg: <https://osf.io/2h4w8/>

Correspondence concerning this article should be addressed to Chung-hong Chan, A5, 6 (section A), 68159 Mannheim, Germany. E-mail: chung-hong.chan@mzes.uni-mannheim.de

Abstract

Comparative approaches such as comparative content analysis are frequently used in journalism studies. The purpose of this paper is to argue that researchers should refrain from using current (frequentist) analytical approaches to analyze data from a comparative study. It is because current approaches cannot account for non-atomicity (nested nature) and non-stochasticity (non-random sampling) of the data. Bayesian multilevel regression is suggested to be a —and probably, the only— valid analytical approach. Using the openly available *useNews* data and the R package *brms*, we demonstrate how to apply the Bayesian approach for the analysis of comparative data. We also address common queries of the approach such as choosing the priors and diagnosing the convergence of the model. Furthermore, we argue that choosing a Bayesian approach also solves the problem of often misinterpreted frequentist confidence intervals.

Keywords: Bayesian inference, multilevel model, comparative communication research, ecological effect

Word count: 7618

Bayesian (comparative) journalism studies

“The comparative approach attempts to reach conclusions beyond single cases and explains differences and similarities between objects of analysis against the backdrop of their contextual conditions.” Esser and Vliegenthart (2017)

Communication research is often comparative, journalism studies are no exception. For example, comparative approach is used in modeling journalistic culture (e.g. Esser & Umbricht, 2013), comparative news value research (e.g. Burggraaff & Trilling, 2017; Wilke, Heimprecht, & Cohen, 2012), news flow research (e.g. Wu, 2000; Grasland, 2019), among others. A typical comparative study in journalism research, otherwise known as comparative content analysis, includes news articles from various outlets and usually also from various countries. Then the included news articles are coded either manually or automatically. Surveys in journalism research are often comparative, e.g. the Worlds of Journalism Study (WJS) (Hanusch & Hanitzsch, 2017).

Although some comparative studies are purely descriptive, most of these studies are aimed to study how contextual conditions such as characteristics of media outlets (e.g. political orientation, online versus offline) and their countries (e.g. post-communist country) influence the content of news articles or behaviours of journalists. For example, some hypotheses are “the coverage of foreign countries by the news is primarily determined by geographical proximity and the status of the covered country” (Wilke et al., 2012, p. 306) and “the degree of opinion-orientation will be highest in newspapers from Polarized Mediterranean systems and lowest in those from Anglo-American systems.” (Esser & Umbricht, 2013) Employing the language from epidemiology (Susser, 1994), the effect on media content and journalistic behaviors in these hypotheses is assumed to be ecological. It assumes a macro contextual factor (e.g. newspapers from Polarized Mediterranean systems) is associated with an outcome at the micro-level (e.g. degree of opinion-orientation), namely, the article-level or journalist-level.

This study of ecological effects on individual behaviors has a long tradition in social sciences. A famous example was Émile Durkheim’s investigation on the ecological factors associated with suicide (Durkheim, 1897). Durkheim found the higher suicide rate among citizens from Scandinavian countries than those from other European countries. Using today’s standard, Durkheim’s analysis is descriptive and this descriptive approach for studying ecological effect is indeed still useful for hypothesis generation in comparative research. An example of hypothesis-generating comparative research is C.-h. Chan et al. (2020), in which the sentiment profiles of terrorism coverage from Muslim- and Christian-majority countries were visualized. Hypothesis-generating comparative study, however, is rare. Most studies have a confirmatory outlook. As we see from the example hypotheses above, most of these studies propose hypotheses to test for ecological effects. In this context usually traditional (frequentist) hypothesis testing approaches were used: Esser and Umbricht (2013) lumped multiple outlets from the same country together and then used univariate analysis of variance to test for the differences in the proportion of opinion-orienting articles across countries. In Wu (2000), multiple stepwise regression was used. These approaches violate the underlying assumptions of the statistical analyses. It highlights the fact that comparative research introduces a feature that researchers usually overlook: media contents are clustered in a multilevel structure. A news article is nested within its media outlet and its media outlet is in turn nested within its country.¹ Such data structure brings two problems: non-atomicity and non-stochasticity.

In our paper we first discuss these two issues before we suggest Bayesian multilevel regression as a standard approach for comparative research in journalism studies. While Bayesian models might be slightly more complicated to estimate in contrast to frequentist models as priors have to be chosen, these models allow researchers to make “principled statements about precision and plausibility” (Morey, Hoekstra, Rouder, Lee, &

¹ Theoretically with a large enough data it can be argued that a news article is nested within authors.

Wagenmakers, 2015, p. 121) and solve the issue of misinterpreted confidence intervals (Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Rinke & Schneider, 2018). Using the openly available *useNews* and WJS datasets and the R package *brms*, we then demonstrate how to apply a Bayesian approach for the analysis of comparative data.

Non-atomicity

“When it comes to regression, multilevel regression deserves to be the default approach” McElreath (2020, p. 356)

It is perhaps well-known what ecological fallacy is. We commit ecological fallacy, when we make inference at micro-level (e.g. Eating more chocolate makes one cleverer) based on macro-level data (e.g. the correlation between countries’ chocolate consumption and countries’ number of Nobel winners). The reverse of ecological fallacy, i.e. making inference at macro-level based on micro-level data, is equally fallacious. This fallacy is called atomistic fallacy (Gnaldi, Tomaselli, & Forcina, 2018).

In comparative research, it is easy to incorrectly assume that macro-level independent variables (e.g. Anglo-American systems) could be analyzed at the same level as the micro-level dependent variable (e.g. the degree of opinion-orientation). The manifestation of this incorrect assumption is to enter macro-level variables as independent variables in multiple regression analysis and regress them against a micro-level dependent variable. Suppose one wants to study a classic news value question of how the probability of covering China changes when the hosting country of the media outlet has a higher trade volume with China. It is related to the news value of “meaningfulness” in Galtung and Ruge (1965). The relationship can be expressed in the following regression equation: let y_i be the dependent variable at the micro-level (coverage of China or not) and x_i be the independent variable at the macro-level (trade volume with China). Suppose there are m news articles where $i = 1, 2, \dots, m$.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (1)$$

This method is the so-called “disaggregation approach”. From an epistemic standpoint, searching for ecological effect by studying the value of slope (β_1) leads to atomistic fallacy (Hox, Moerbeek, & Van de Schoot, 2017). From a statistical standpoint, this approach violates the underlying independence assumption. Almost all frequentist statistical tests assume that the observations are independent of each other: y_i given x_i are independently identically distributed (i.i.d.). Articles from the same media outlet are hardly i.i.d.: they are subjected to being edited by the same editorial team, to being written by the same group of journalists exposed to the same journalistic culture and paid for by the same employer. Also, the way communication researchers collect their media content resembles cluster sampling rather than random sampling: one usually starts from a specific media outlet and then collects all relevant articles from it. The choice of media outlets or countries are almost always not randomly selected.² This sampling approach makes the independence assumption even more fragile.

Previous simulation studies have shown that ignoring this dependence can lead to false-positive associations (e.g. Clarke, 2008; Chen, 2012). The problem is more severe when the dependent variable is a discrete binary variable (Clarke, 2008), a common feature of content analysis.³

² If countries were randomly selected, we shall see more African countries in comparative works. It is because 1/4, or 54 out of 195, UN countries are in Africa.

³ The non-atomicity argument has been in standard statistics textbooks for 40+ years. One could question the necessity of another repetition in this paper. The non-atomicity argument might be repetitious but we believe it is still necessary, not least because multilevel modeling –unlike what McElreath (2020, p. 356) said– is still not the default regression approach in communication research. For example, as recent as in 2020 some communication researchers apply the disaggregation approach by justifying the intraclass

There are two solutions to this. The first solution is to aggregate the micro-level dependent variable across a macro-level independent variable. Suppose there are n countries in the above example. One can aggregate the total number of articles covering China as z_k for the k -th country, where $k = 1, \dots, n$. Then, we can do a count-based regression with a regression equation like so:

$$\log z_k = \beta_0 + \beta_1 x_k + \epsilon_k \quad (2)$$

Using this aggregation method, the unit of analysis effectively switches from article to country. This method is useful when x is the only independent variable, akin Esser and Umbricht (2013). It is not technically committing the atomistic fallacy, when one uses the value of slope as the evidence for an ecological effect. However, this method still has important drawbacks. First, it cannot be used for data with more than two levels. In Esser and Umbricht (2013), for example, multiple media outlets from the same country were lumped together and in effect, assumed to be homogeneous and the information about media outlets was discarded. Second, this method discards a massive amount of information. It is better for analysis of a dependent micro-level variable with a reasonable aggregation function (e.g. counting a binary variable). For numerical variables, aggregation functions such as taking a mean or median cannot capture the spread of the micro-level variable (Bryk & Raudenbush, 1988). Also, the effective sample size is reduced from the number of articles (m) to the number of countries (n). Nonetheless, this aggregation method, although inflexible, is still useful when the number of groups (e.g. n) is large. It is

correlation between groups is small and the impact of ignoring the multilevel structure is also small (a finding in Chen, 2012), see an example in Zhu and Fu (2020 footnote 5). Even a multilevel modeling textbook advocates this justification (Bickel, 2007). We reject this justification because it violates the independence assumption and thus commits atomistic fallacy. A fallacious result with less error is still fallacious. See also Nezlek (2008) for a criticism of such justification.

also useful to collapse micro-levels (e.g. article-level) that are not useful in answering one's research questions.

Another solution is to use the multilevel model (linear mixed model, or hierarchical model). In a multilevel model, the effect on the micro-level dependent variable (y) is modeled with equations at different levels. Using the above example, y_{ik} denotes whether the i -th article in the j -th country is covering China; x_k denotes the trade volume with China of the k -th country where the media outlet of i -th article is located.

$$y_{ik} = \beta_0 + \epsilon_{ik} \tag{3}$$

$$\beta_0 = \gamma_{00} + \gamma_{01}x_k + \mu_{0k} \tag{4}$$

In these equations, γ_{00} is the average slope, while μ_{0k} is group-dependent deviations of the slope from the average. It is usually set as having a normal distribution with a variance τ_{00} , i.e. $\mu_{0k} \sim \mathcal{N}(0, \tau_{00})$. Instead of a single value, the regression coefficient γ_{00} is assumed to be a distribution of values depending on a macro-level group. It addresses the problem of clustering of articles by macro-level variables. This model is called the varying-intercept model and is used frequently in social science research. We can then study the magnitude of γ_{01} to determine the ecological effect.

The benefit of using multilevel modeling lies in its flexibility in handling multi-level data. Suppose we also want to consider the clustering of articles around o different media outlets in the above example and j -th media outlet is the media outlet of the i -th article, where $j = 1, \dots, o$. The multilevel regression equations are rewritten as:

$$y_{ijk} = \pi_0 + \epsilon_{ijk} \quad (5)$$

$$\pi_0 = \beta_{00} + \mu_{0j} \quad (6)$$

$$\beta_{00} = \gamma_{000} + \mu_{00k} + \gamma_{001}x_k \quad (7)$$

This flexibility is demonstrated in the study by Rinke (2016). He studied the likelihood of opinion justification in 1559 utterances nested in 329 news items, which were in turn nested in 101 news broadcasts. Multilevel logistic regression was used to model the natural three-level hierarchy of his data.

Multilevel modeling is the way to go for analyzing non-atomic data from comparative research. However, the conventional approach for statistical inference makes it not optimal. The conventional approach for statistical inference in multilevel modeling is maximum likelihood estimation (MLE). Stegmueller (2013) demonstrates that MLE is associated with shrinkage (reduction of standard error, i.e. more false positives) and the shrinkage is more severe when the number of macro-level units (e.g. countries) is small. Stegmueller (2013) proposes to use Bayesian analysis as a robust alternative (see counterarguments from Elff, Heisig, Schaeffer, & Shikano, 2020; Bryan & Jenkins, 2015). Although less restrictive methods such as restricted maximum likelihood (REML) have been demonstrated to remediate the shrinkage issue of MLE (Elff et al., 2020), we still agree with Stegmueller’s proposal for theoretical reasons. Our argument is more in line with Western and Jackman (1994).

Non-stochasticity

*“If Czech history **could** be repeated, we should of course find it desirable to test the other possibility each time and compare the results. Without such an experiment, all considerations of this kind remain a game of hypotheses.”* (Kundera, 1984, emphasis added)

Before diving into our theoretical reasoning, it is important to revisit what frequentist inference, the current default but often misunderstood (Rinke & Schneider, 2018) mode of inference in communication science, is. Under the frequentist framework, each experiment is assumed to be one of infinite independent, **repeatable** experiments on randomly drawn samples from a population. Random experiments are assumed to be repeated arbitrarily often. Based on this assumption and with just one experiment from the current study, we make an estimation about the population. The discrepancy between the estimation from that one experiment and the actual value of the population is due to sampling error alone, i.e. which subjects were randomly sampled from the population. Randomized surveys, for example, are assumed to be repeatable through repeated random sampling of the population. Suppose we replicated the same survey for 100 times and we would obtain a slightly different sample every time. We then calculated the 90% confidence interval of the mean for each of these 100 surveys. We should anticipate that more or less 90 out of these 100 confidence intervals would include the true mean of the population. We cannot say for sure exactly 90 out of these 100 confidence intervals would include the true mean of the population because repeated random sampling is indeed random and the process is **stochastic**. However, we can say 90 is more probable than 0 or 100.

Comparative content analytic studies in journalism studies, unlike randomized surveys or randomized media experiments, often collect all available data. In contrast to sampling, these studies are actually doing census of all data. There is no way to get more data unless the scope of these studies is changed (Berk, Western, & Weiss, 1995). It is especially true for modern large-N studies using automated content analytic techniques. Burggraaff and Trilling (2017), for instance, “collected all available news items from a selection of major Dutch news outlets, both online and print” (p. 6) and that amounted to 762,095 articles from 9 outlets in the period of 2014–2015.

Census-style studies are not repeatable and thus they generate non-stochastic data.

In other words, data from comparative journalism studies are fundamentally irrelevant for frequentist inference (Western & Jackman, 1994). P-values and confidence intervals generated do not have the same meaning as those from repeatable studies. According to Western and Jackman (1994, p. 413), these values from non-stochastic data “lack meaning even as abstract propositions”.

A common counterargument to this is the classic one from Deming and Stephan (1941, p. 45), who suggested that “[a]s a basis for scientific generalizations and decisions for action, a census is only a sample.” This notion assumes a census of a population can be used to make inference on a theoretical device called *superpopulation* which “theoretically could exist, may have existed, or may exist in the future” (Gibbs, Shafer, & Miles, 2015, p. 3). In other words, a finite census-as-a-sample census is assumed to be a “representative sample” of an infinite superpopulation. This counterargument could be useful but unlike a regular sample, whose representativeness can be assessed, we can never assess the representativeness of the census with respect to the theoretical superpopulation. Echoing the quote at the beginning of this paragraph, there is no way to tell if the current Czech history is representative of all possible Czech histories in the multiple parallel universes.

Instead of invoking the theoretical device of superpopulation, we follow the arguments from Western and Jackman (1994) and Stegmueller (2013) for comparative studies: Bayesian inference should be used for analyzing data from comparative research in our field. Choosing a Bayesian approach also solves the problem of misinterpretation of confidence intervals (Hoekstra et al., 2014; Rinke & Schneider, 2018). Switching to a Bayesian approach addresses this issue as Bayesian “credible intervals support an interpretation of probability in terms of plausibility” (Morey et al., 2015, p. 120).

Before we move on to the next section, it is important to point that **non-stochastic data is not an essential condition for applying the Bayesian approach**. The same approach is equally applicable to both non-stochastic and stochastic comparative datasets.

Therefore, one can analyse both randomized comparative surveys (such as WJS) and comparative content analytic studies with the same Bayesian analytic approach.

Bayesian analysis

“PROBABILITY DOES NOT EXIST.” Bruno de Finetti

What is the probability for this paper being accepted by *Digital Journalism*? Under the frequentist framework (and if *Digital Journalism* were accepting papers stochastically), we can only find this out by repeatedly submitting this paper to *Digital Journalism*, say for 100 times, and then count the frequency of acceptance in these repeated submissions. It is indeed impractical as well as inhumane to the editorial team of *Digital Journalism*. Instead, we assert before submission that this paper has a probability of 24% for being accepted. That is the published acceptance rate from *Digital Journalism*’s website. After this paper is submitted and is not desk rejected, the probability might be around 24% to 30%. After a month of waiting and our confidence is shaken a little, the probability might decrease to 10% to 20%. After the paper is favorably reviewed by two reviewers and an R&R is invited, the probability might increase to 40% to 60%. If you see this paper under the “Latest articles” section of *Digital Journalism*’s website, then the probability is beyond doubt 100% (or 0% if we submit this paper again).

Without repeated experiment, these probabilities quantify our certainty on how likely this paper is being accepted, given the current available data . We revise our old beliefs (or **prior**, $p(\theta)$) with the new data (X) and form our revised belief (or **posterior**, $p(\theta|X)$). This can be summarized in the following equation (Gelman et al., 2020):

$$P(\theta|X) \propto P(\theta)P(X|\theta) \tag{8}$$

The $P(X|\theta)$ part is called likelihood function. In the *Digital Journalism* example, the

likelihood function is based on rough rules from our experience and thus is not systematic.⁴ In actual analysis, we need to derive such a likelihood function based on the available data using methods such as Markov Chain Monte Carlo (MCMC). Nonetheless, the above equation indicates that there are only three ingredients in any Bayesian analysis: 1) data (X), 2) a method to derive the likelihood function $P(X|\theta)$ from the data, and 3) prior, $P(\theta)$.

We skip elaborating part 1, because some, but certainly not all, researchers have the means to amass a massive amount. The mechanism of how part 2 can be used to derive the likelihood function is beyond the scope of this paper. The current standard method, MCMC, is much more computationally intensive and has only been possible on desktop computers since the mid-1990s (Robert & Casella, 2011). And only in the last few years that can be done efficiently (Hoffman & Gelman, 2014), thanks to the probabilistic programming language *Stan* (Gelman, Lee, & Guo, 2015). R interfaces to Stan, such as *brms* (Bürkner, 2017), make Bayesian analysis even more accessible to R users. But still, it is important to be mindful that Bayesian analysis is much more computational intensive than methods such as MLE. Our benchmark suggests that Bayesian analysis needs at least 100 times more running time than MLE.

Part 3 (Prior) is arguably the most controversial part of Bayesian analysis. In the *Digital Journalism* example, we can select a reasonable prior (or **informative** prior) of 24% from published information. Finding previous studies for an informative prior is the logical first thing to do. Keating and Totzkay (2019) show that one in every seven communication research papers published in major communication journals was a form of replication attempt. For these one seventh of communication research, there should be previous studies available to base one's informative prior non-controversially.

⁴ As a side note: we have actually performed the experiment by submitting this paper to *Digital Journalism*. But for pragmatic reasons we cannot report the posterior.

For the other six sevenths, one might not have any information to set an informative prior. One option is to consult experts or to make an educated guess. Expert elicitation is a way to probe how the experts in the field think about the hypotheses. A standardized protocol for expert elicitation is available (Hanea et al., 2017) and there are many software tools available to facilitate the process.⁵

But one person’s expert opinion could be another person’s wishful thinking. And this perceived subjectivity of specifying priors by experts’ judgement attracts wide-spread criticism from both statisticians (e.g. Efron, 1986) and social scientists (Bryan & Jenkins, 2015; Elff et al., 2020).

Undoubtedly, setting prior is consequential to the analysis (Van Dongen, 2006). But the influence from priors is greatly weakened, when the data is getting bigger. Other than expert elicitation, another non-controversial way to specify priors — at least in our opinion — is to use a weakly informative prior (Lemoine, 2019). In this way, one specifies only the possible range of the posterior.⁶ Some so-called “default weakly informative priors”, e.g. $t(1, 0, 0.25)$, have been suggested for typical regression models (Gelman, Jakulin, Pittau, & Su, 2008). The documentation of Stan also provides information on how to choose an appropriate prior (<https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>).

Bayesian communication research

“There is a general challenge to a prescription of more widespread use of Bayesian methods for multilevel modeling, and that is that such methods require statistical expertise

⁵ Some examples are the web-based MATCH Uncertainty Elicitation Tool (<http://optics.eee.nottingham.ac.uk/match/uncertainty.php>) and the R package SHELF.

⁶ One could also use a completely *noninformative* prior such as $\mathcal{U}(-\infty, \infty)$. It is, however, no longer advised.

beyond that of most applied social science researchers, as well as specialist software (or software with which such researchers are unfamiliar)." Bryan and Jenkins (2015, p. 19)

Bayesian analysis is still a minority statistical method in social sciences. It is reasonable to say that Bayesian analysis is extremely rare in our field. Although some communication researchers adopted the method (e.g. Kovic, Rauchfleisch, Metag, Caspar, & Szenogrady, 2017; C. Chan & Grill, 2020) these studies are not comparative. As far as we know, the only available comparative research in journalism studies that Bayesian multilevel regression was used for studying ecological effect is de Leeuw, Azrout, Rekker, and Van Spanje (2020). We believe comparative communication research fits the use case of Bayesian multilevel regression analysis. Although there is another methodological introduction to Bayesian analysis aiming at communication researchers (Konijn, van de Schoot, Winter, & Ferguson, 2015), the main purpose of this paper is to demonstrate how to actually do the analysis using the R package *brms* (Bürkner, 2017). As the interface of *brms* is almost the same as *lme4* (Bates, Mächler, Bolker, & Walker, 2015, another R package for fitting multilevel models using MLE), *lme4* users might find *brms* extremely familiar. The above quoted challenge about the prescription of more widespread use of Bayesian methods for multilevel modeling is no longer an issue.

The two examples below show how the Bayesian multilevel regression analysis can be deployed to the common archtypes of comparative journalism studies. The first example is an analysis of a comparative, non-stochastic content analytic dataset. The second example is an analysis of a comparative, stochastic survey dataset.

Example 1: useNews

We used the useNews dataset (Puschmann & Haim, 2020) to demonstrate how to perform a typical Bayesian multilevel regression analysis on a nonstochastic comparative dataset. The analysis was guided by a classic question in news value research: Does the

distance between China and the host country of a media outlet increase the frequency of China coverage? In other words, we studied the ecological effect of the distance from China on the frequency of China coverage. Several distance measures were used to formulate our pre-registered hypotheses: physical distance between two capitals, trade volume between two countries, and cultural distance between two societies.

For demonstrative purposes, our analysis was separated into two levels: outlet-level analysis and article-level analysis. The outlet-level analysis is an aggregated version of article-level analysis. It is used to demonstrate the flexibility of Bayesian analysis to handle two-level data with a very small data size. The article-level analysis is used to demonstrate the analysis of a massive three-level dataset.

Data

The useNews is an openly available dataset (Puschmann & Haim, 2020) that includes 2019 to 2020 media content data from an array of worldwide media outlets . The media content data was collected from the MIT MediaCloud and made available as document-term matrices.

We used only the data from 2019 as the baseline. It is important to note that the Hong Kong protests, the 30th anniversary of the Tiananmen Square Protests and the 70th anniversary of the communist regime were in that year. Also, it was at the zenith of the US-China trade war.

We excluded media outlets that contributed <1,000 articles in that year. This threshold allowed us to retain media outlets that the frequency of China coverage can be reliably estimated. In total, 61 media outlets from Spain ($n = 11$), Romania ($n = 11$), US ($n = 7$), Austria ($n = 6$), Germany ($n = 5$), Norway ($n = 5$), Australia ($n = 4$), Brazil ($n = 4$), UK ($n = 4$), South Korea ($n = 2$), and the Netherlands ($n = 2$) were included. These 61 media outlets contributed 1,525,871 articles.

In the outlet-level analysis, the data has 61 rows and each row represents a media outlet. The data contains a count of articles covering China (z), total number of articles (n), country of the outlet(k), and distance measures (x).

In the article-level analysis, the data has 1,525,871 rows and each row represents an article. The data contains a binary dependent variable of China coverage (y), media outlet(j), country of the outlet(k), and distance measures (x).

Coverage of China

A dictionary-based approach was used. The seed English and German dictionaries from the R package `newsmap` (Watanabe, 2017) were used as the basis. The seed dictionaries contain words about Chinese, China, Beijing and Shanghai. We developed further the Spanish, Romanian, Korean, Portuguese, Norwegian, and Dutch dictionaries.⁷

⁷ One might argue our seed dictionaries are not validated. We deem that the criterion validity of our dictionary should be built-in. Coverage of China, by definition, should mention words about China. We also maintain that it is difficult to conduct a human validation of the seed dictionaries because the useNews dataset (Puschmann & Haim, 2020) doesn't provide access to the full-length articles. Nonetheless, we attempted to conduct such validation by randomly selecting 10 matching and 10 non-matching items for each language. In total, 160 items were selected. A coder with no knowledge of the ground truth (matching or not) annotated the full-text articles by following the URLs in the meta data provided in the useNews dataset (Puschmann & Haim, 2020). For some languages, the coder used Google Translate to translate the full text article to English. In total, only 138 URLs were still accessible. Comparing the ground truth and the human coding, 62 and 70 are true-positive and true-negative cases. There are also 5 false-negative and 1 false-positive cases. There are several reasons not related to our seed dictionaries that could lead to the 6 misclassified cases. The false-positive case is Korean and that could be due to incorrect tokenization at the MIT Media Cloud's end. The 5 false-negative cases are in Norwegian, Portuguese, and Romanian. The content on the website might have changed since MIT Media Cloud collected the data. For example, the ever-changing content in the sidebar might be counted as "main content" and recorded in the DTM. Despite all these caveats, our seed dictionaries have a precision and recall of 98.4% and 92.5% respectively. For the purpose of demonstrating regression analysis, this level of accuracy is more than sufficient.

The dictionaries were applied to the 61 document-term matrices (format of the provided dataset) of all included media outlets. An article is classified as China coverage when at least one dictionary match is detected.

Distance measures

Several distance measures were used as the independent variable to test our central research questions: physical distance, trade volume, and cultural distance. Physical distance is calculated as the Great-circle distance between Beijing and the capital of the hosting city of the media outlet (e.g. Bucharest). The volumes of import to and export from China with the host country of a media outlet was extracted from the 2019 edition of China Statistical Yearbook (<http://www.stats.gov.cn/tjsj/ndsj/2019/indexeh.htm>). Cultural distance from China to the host country of a media outlet was extracted from culturaldistance.com, which was calculated using the data from World Value Survey with the algorithm from Bell, Richerson, and McElreath (2009). All of these distance measures were log-transformed, because previous studies showed the relationship between news coverage and these distance measures is not linear (Grasland, 2019; Wu, 2000).

Modeling

For all analyses, we tried 3 different approaches: 1) Disaggregation approach (ignoring the multilevel structure); 2) Multilevel regression using MLE; and 3) Bayesian multilevel regression. It is important to remind our readers that the results from the three are likely to be similar, but it is not evidence for the disaggregation approach and the MLE approach is a replacement of the Bayesian approach (Morey et al., 2015). It is because the disaggregation approach commits atomistic fallacy and the MLE approach ignores the non-stochastic nature of the data.

Outlet-level analysis. We used negative binomial regression for this analysis. Media outlets are nested in countries (e.g. *Der Standard* and *Österreichischer Rundfunk* are nested in Austria). Thus, a country-based variance is added (μ_{0k}). We also added

$\log n_j$ as an offset value. An offset value is a term that does not have the associated regression coefficient at the right-hand side of the regression equation (2). Effectively, we modeled the *rate* of China coverage. Applying the same notation under the introduction section, the multilevel regression equations are:

$$\log z_{jk} = \gamma_{00} + \mu_{0k} + \gamma_{01}x_k + \log n_j + \epsilon_{jk} \quad (9)$$

$$\log z_{jk} - \log n_j = \gamma_{00} + \mu_{0k} + \gamma_{01}x_k + \epsilon_{jk} \quad (10)$$

$$\log \frac{z_{jk}}{n_j} = \gamma_{00} + \mu_{0k} + \gamma_{01}x_k + \epsilon_{jk} \quad (11)$$

The estimand of interest, γ_{01} , is interpreted as the average unit change in the log rate of China coverage, $\log \frac{z_{jk}}{n_j}$, for each unit change in the distance measure of the outlet k .

In lme4 and brms, the above equation is expressed as such:

```
require(brms)
require(lme4)

# brms
brm(z ~ offset(log(n)) + (1|k)+log(x),
    family = negbinomial(), data = outlet_data)

# lme4
glmer.nb(z ~ offset(log(n)) + (1|k) + log(x), data = outlet_data)
```

In the disaggregation case, the multilevel structure is ignored and the regression equation becomes:

$$\log \frac{z_{jk}}{n_j} = \beta_0 + \beta_1 x_k + \epsilon_j \quad (12)$$

In MASS, it is expressed as such:

```
require(MASS)
glm.nb(z ~ offset(log(n)) + log(x), data = outlet_data)
```

We named the models with different independent variables as model A1-4 respectively (A1: log import volume; A2: log export volume; A3: log physical distance; A4: log cultural distance).

Article-level analysis. We used logistic regression for this analysis. Due to the computational time needed to conduct the Bayesian regression in this case (it took 4 days on a regular computer), we only did the model with import volume as the independent variable.

In this analysis, an article is nested in its media outlet. The media outlet is in turn nested in its country. Two variance terms (μ_{0j} and μ_{00k}) are needed. The multilevel regression equations are:

$$p_{ijk} = Pr(y_{ijk} = 1) \quad (13)$$

$$L_{ijk} = \log \frac{p_{ijk}}{1 - p_{ijk}} \quad (14)$$

$$L_{ijk} = \pi_0 + \epsilon_{ijk} \quad (15)$$

$$\pi_0 = \beta_{00} + \mu_{0j} \quad (16)$$

$$\beta_{00} = \gamma_{000} + \mu_{00k} + \gamma_{001}x_k \quad (17)$$

Or in lme4/brms code:

```
brm(y~(1|k/j)+log(x), family = bernoulli(), data = article_data)
glmmr(y~(1|k/j)+log(x), family = binomial(), data = article_data)
```

In the disaggregation case, the multilevel structure is ignored and the regression equations become:

$$p_{ijk} = Pr(y_{ijk} = 1) \quad (18)$$

$$L_{ijk} = \log \frac{p_{ijk}}{1 - p_{ijk}} \quad (19)$$

$$L_{ijk} = \beta_0 + \beta_1 x_k + \epsilon_{ijk} \quad (20)$$

In MASS, it is expressed as such:

```
glm(y ~ log(x), family = binomial(), data = article_data)
```

We named this model as model B1.

Priors

For the outlet-level analysis, there are four unknown parameters to be estimated: γ_{00} , μ_{0k} , γ_{01} and the negative binomial shape parameter ϕ . For the article-level analysis, the four unknown parameters are γ_{000} , μ_{00k} , γ_{001} and μ_{0j} . The idea of choosing priors is to select a probable probability distribution for each of these unknown parameters.

Usually, the first step for choosing priors is to review previous studies and look for possible values to be used as our informative priors. Although Wu (2000) is a possible reference point, we select not to use this because China was not studied. Also, the modeling technique was quite different from the current one.

In this analysis, we used weakly informative priors suggested by Lemoine (2019).

For regression coefficients, e.g. γ_{00} , we used a normal distribution of $\mathcal{N}(0, 1)$. For variance terms, e.g. μ_{0k} , a student t-distribution of $t(3, 0, 2.5)$ was used. A gamma distribution of $\text{Gamma}(0.01, 0.01)$ was used for the shape parameter ϕ .

All, except the prior for regression coefficients, are default priors suggested by *brms*. The prior can be set with the following R code.

```
weaklyinformative_prior <- c(prior_string("normal(0, 1)",
                                         class = "b"),
                             prior_string("normal(0, 1)",
                                         class = "Intercept"))
# The sample_prior argument is optional. But useful for further analysis.
import_brms <- brm(z~offset(log(n))+(1|k)+log(x_import),
                  data = outlet_data, family = negbinomial(),
                  prior = weaklyinformative_prior, sample_prior = TRUE)
import_brms
```

Other parameters. Other parameters for Bayesian multilevel regression analysis control how the MCMC should be performed. Three parameters are important: number of iterations, number of chains and *adapt_delta*. They are not some magic numbers that you can plug and play. However, it is in general safe to use the default values, i.e. number of iterations = 2000, number of chains = 4 and *adapt_delta* = 0.8, unless MCMC shows evidence of nonconvergence. *brms* is smart enough to give suggestions on how to adjust these three values. Nonetheless, we provide a short guide on how to diagnose convergence.

Regression results

Table 1 shows a summary of all models from outlet-level analysis using three different ways of modeling. The Bayesian multilevel regression gave the *posterior distribution*, $P(\theta|X)$, of the regression coefficient (γ_{001}). It is a distribution and therefore we need ways to display both the central tendency and spread of the distribution. By default, *brms* displays the mean and the 95% high density interval (HDI).⁸ These two values represent

⁸ HDI is one representation of the Bayesian credible interval. It is important to note that, like confidence intervals, there should be no “correct” choice of the level of credibility but some established conventions.

the point and interval estimates of the regression coefficient.

Consistently, point estimates are remarkably similar. Bayesian multilevel regression gave a wider 95% HDI than confidence intervals from frequentist methods. It is similar to previous studies (Elff et al., 2020; Stegmueller, 2013). The results indicate that only a higher export or import volume with China and the host country of the media outlet is associated with more reporting of China. While the frequentist confidence intervals and the Bayesian credible interval seem to be similar, they are from a philosophical point of view totally different and lead to different forms of inference. The Bayesian approach allows us to say that *given the observed data, the true estimate of log import volume has 95% probability of falling within 0.18 and 0.33*. In contrast, the frequentist confidence interval only allows us to say that *when computing a confidence interval from the same type of data in repeated studies (if possible), 95% of the confidence intervals will include the true estimate of log import volume*. One of the misconceptions of frequentist confidence intervals is that they can be interpreted in a Bayesian way as described above (Morey et al., 2015). This example illustrates that a Bayesian approach allows us to interpret the results in a more intuitive way.

We choose, like many published works (e.g. Stegmueller, 2013; C. Chan & Grill, 2020; Kovic et al., 2017), the default value of 95% to maintain “compatibility” with the established convention of reporting 95% confidence interval. However, there is a general concern about this credibility level as being unstable for Bayesian analysis (Kruschke, 2014). Some Bayesians, notably, Andrew Gelman, report very liberal 50% credible intervals. McElreath (2020) selects 89% to show the arbitrariness of the practice. de Leeuw et al. (2020) select 90%.

Table 1

Point and interval estimates from outlet-level analysis using three different analytic strategies: Disaggregation approach, multilevel modeling using MLE and Bayesian multilevel modeling

Model	Disaggregate (β_1 , 95% CI)	MLE (γ_{01} , 95% CI)	Bayesian (γ_{01} , 95% HDI)
Model A1: Log import volume	0.25 (0.18, 0.33)	0.25 (0.18, 0.33)	0.25 (0.18, 0.33)
Model A2: Log export volume	0.22 (0.14, 0.30)	0.22 (0.13, 0.31)	0.22 (0.11, 0.33)
Model A3: Log Physical distance	-0.02 (-0.34, 0.24)	-0.10 (-0.55, 0.31)	-0.12 (-0.63, 0.39)
Model A4: Log Cultural distance	-0.45 (-1.26, 0.34)	-0.60 (-1.80, 0.55)	-0.62 (-2.12, 0.78)

Model B1 confirms the above again (Table 2). Therefore, both article- and outlet-level analyses arrive at the same conclusion.

Table 2

Bayesian multilevel logistic regression analyzing the article-level likelihood of China coverage in relation to log import volume of the outlet's country

Parameter	Estimate	SD	95 %HDI
γ_{000} (Intercept)	-7.06	0.81	(-8.61, -5.44)
γ_{001} (Log import volume)	0.27	0.06	(0.16, 0.38)

Diagnosis of convergence. It is important to confirm the model is converged. We demonstrate how to confirm this with model A1.

The first step is Gelman-Rubin convergence diagnostic (Brooks & Gelman, 1998; Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2020). *brms* displays this in the form of \widehat{R} . In general, $\widehat{R} \approx 1$. The model might not have converged, when $\widehat{R} \geq 1.01$ (Vehtari et al., 2020). The *summary()* function of *brms* displays the \widehat{R} for each of the estimated unknown parameters. All parameter estimates have a $\widehat{R} = 1.0$.

The second step is the MCMC trace plot (can be obtained with the function *plot()*). The traceplot should display the so-called “fuzzy caterpillar” pattern (Figure 1). There should be no breakage.

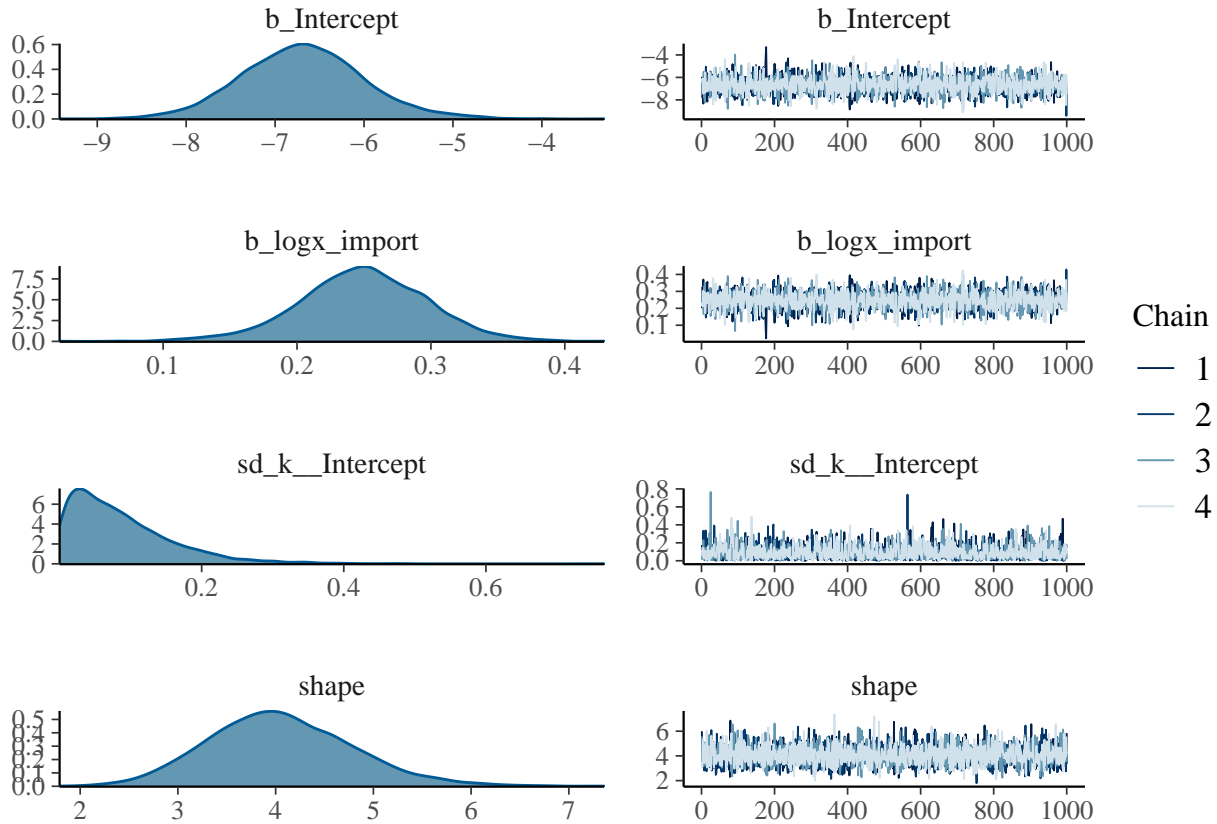


Figure 1. The right panel in the MCMC trace plot from *plot()* function showing the fuzzy caterpillar pattern.

An additional check is to study the autocorrelation of all chains. The correlogram should display low serial correlation (successive high correlation). The function *mcmc_plot* can be used to plot this (Figure 2).

If the diagnostics show evidence of nonconvergence, it is suggested to increase the number of iterations and number of chains.

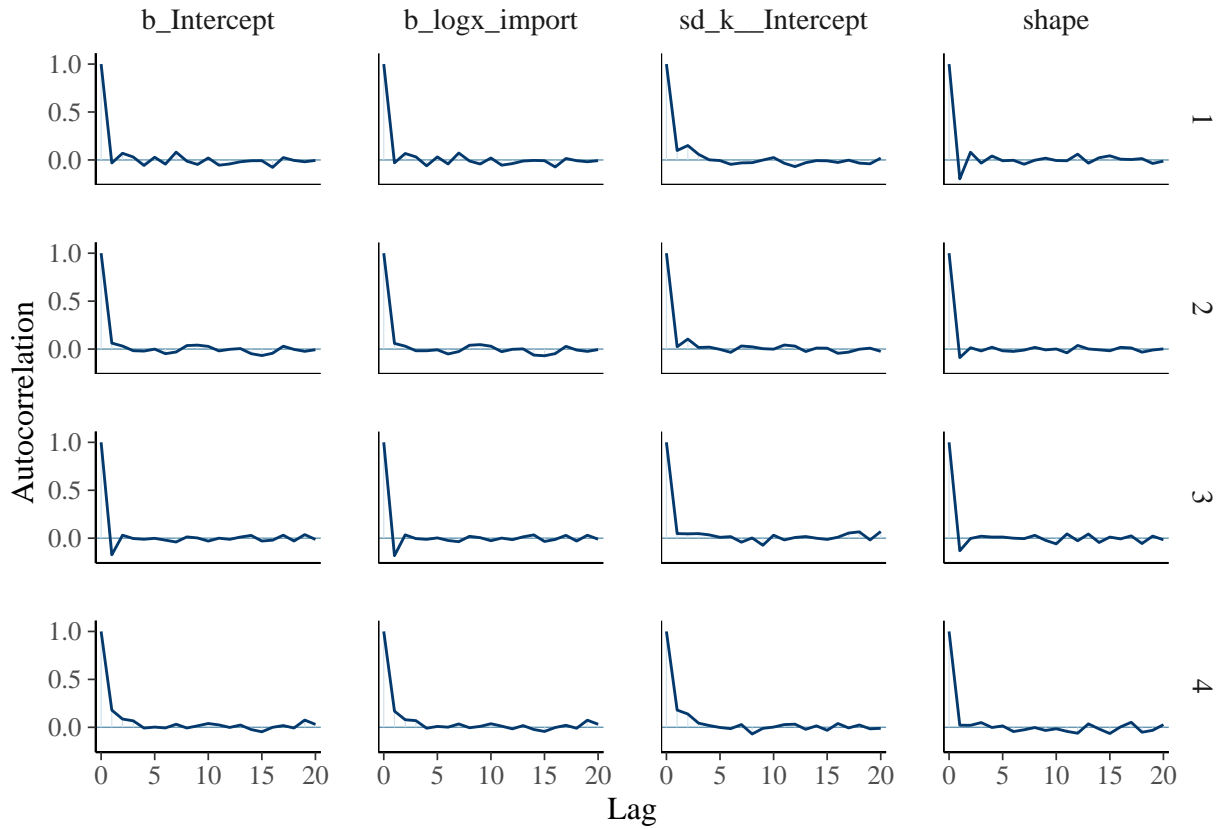


Figure 2. Correlogram of four chains showing no significant higher-order lag

Further diagnostics. There are several diagnostic checks that we can do to make sure our model is valid.

- Increasing number of iterations

The first test is to increase the number of iterations from 2000 to a higher value, e.g. 4000, and see if the result is consistent with the initial analysis. This check makes sure our result is not transient due to short iteration time.

Increasing the number of iterations did change a bit of the estimates, but only at the second decimal place. It shows that the results are consistent between two analyses.

- Using another weakly informative prior

We can also use another weakly informative prior to see how our initial choice influences the result. We tried another weakly informative prior for the regression coefficients, $t(1, 0, 0.25)$ (Gelman et al., 2008). Similarly, changing prior only impacts the second decimal place of the estimates.

- Posterior predictive checks

We can use posterior predictive checks to study how our model fits our data. The gist of the method is to use the fitted model to generate some simulated data (y_{rep}) and compare them with the original data (y). A good model should display a similar probability distribution of y_{rep} and y . It can be done with the function `pp_check()` (Figure 3).

```
## Warning: Argument 'nsamples' is deprecated. Please use argument 'ndraws'
## instead.
```

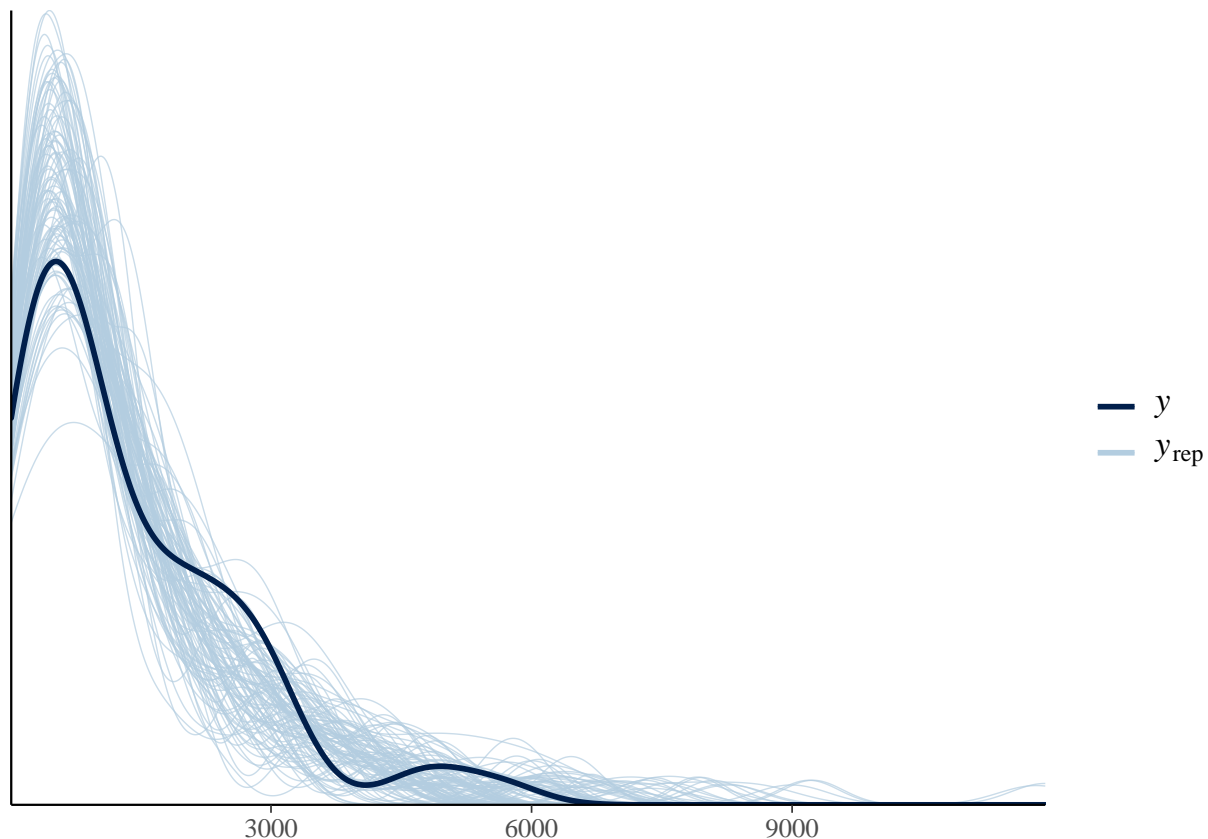


Figure 3. Posterior predictive checks with 100 sets of simulated data

Extensions

We propose several extensions to the above barebone analysis. The analyses in these extensions are not preregistered.

Extension #1: Hypothesis testing. Under the Bayesian framework, we can test hypotheses about the parameter estimates (Shikano, 2019). These hypotheses are similar to the null hypothesis in the Neyman-Pearson’s sense. For example, we want to test the two-sided hypothesis of $\gamma_{01} = 0$ for Model A1. For this, we need to calculate the posterior probability of this hypothesis: $P(\gamma_{01} = 0|X)$. The function *hypothesis()* can be used to test such hypothesis.

```
hypothesis(import_brms, "logx_import = 0")
```

The posterior probability of this hypothesis is 0. Therefore, we can safely reject the hypothesis and conclude that $\gamma_{01} \neq 0$, i.e. log import volume of the outlet’s country is associated with the frequency of China coverage at the outlet level.

While accepting a null hypothesis is not a feasible option in a frequentist framework, as such a framework has an “inability to gain evidence for the null” (Rouder, Speckman, Sun, Morey, & Iverson, 2009, p. 226) hypothesis, it is practically possible to accept a null hypothesis in a Bayesian framework (Kruschke, 2018). Bayes factor or the region of practical equivalence (ROPE) could be used to decide to accept the null hypothesis. Kruschke (2018), for example, suggests using $\sigma = \pm 0.1$ as half of small effect size to establish a ROPE. However, there is no fixed rule on how to define a ROPE. If now 95% of the HDI for an estimate falls within the ROPE, researchers could accept the null hypothesis for practical purposes. A practical application could be with manual content analysis or survey research in communication research if some variables are challenging and costly to measure. In a frequentist framework, failing to reject the null hypothesis could wrongly inform researchers to collect more data in subsequent studies to discern between

insufficient power⁹ or the null is true. Instead, ROPE could be used to decide whether, in future studies, more data should be collected to retest the same hypothesis. When ROPE provides a strong signal that the null hypothesis is true, it is no longer worthy to collect more of the costly data just to boost the statistical power.

The package `bayestestR` (Makowski, Ben-Shachar, & Lüdtke, 2019) can be used to calculate and visualize the ROPE.

```
library(bayestestR)
rope(import_brms)
plot(rope(import_brms))
```

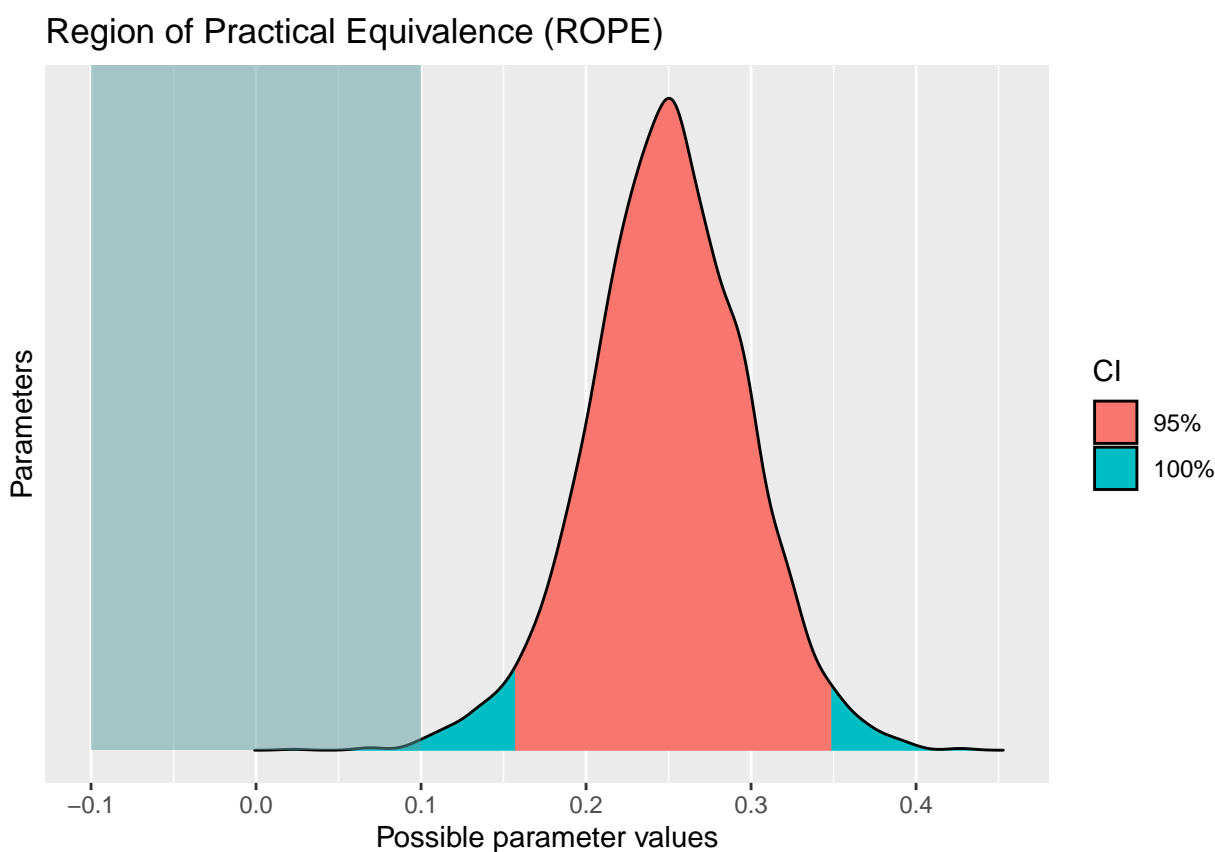


Figure 4. The Region of Practical Equivalence and the 95% High Density Interval

⁹ Statistical power is a frequentist concept specifying the long run probability of correctly rejecting the null hypothesis. It is usually not relevant for Bayesian inference.

The proportion of 95% HDI falling inside the ROPE is 0%. Figure 4 also displays that the 95% HDI (the pink region) have no overlap with the ROPE (the shaded region). Using the decision rules by Kruschke (2018), there are 3 possible scenarios: reject the null (when most of the 95% HDI falls outside the ROPE), accept the null (when most of the 95% HDI falls inside the ROPE) and undecided (when it can't be decided). In the current situation, it is very clear that the 95% HDI falls outside of the ROPE as there is 0% overlap. Similar to the result from *hypothesis()*, there is a practical effect and it is correct to reject the null hypothesis.

However, in another situation when we have a simulated independent variable *noise* that is basically random noise.

```
outlet_data$noise <- rnorm(nrow(outlet_data))

noise_brms <- brm(z~offset(log(n))+(1|k)+noise,
                  data = outlet_data, family = negbinomial(),
                  control = list(adapt_delta = 0.99),
                  prior = weaklyinformative_prior,
                  sample_prior = TRUE)

rope(noise_brms)
```

In this situation, the proportion of 95% HDI falls inside the ROPE is 68%. The proportion is quite high but it is not decisive. Kruschke (2018) suggests accepting the null when more than 95% of the 95% HDI falls inside the ROPE. Accordingly, the current situation of 68% is “undecided”. It might be worthwhile to retest the hypothesis with a larger amount of data.

One should note that the decision rules by Kruschke (2018) depend on many methodological decisions, e.g. how to determine the ROPE, how wide should the ROPE be,

and which level of HDI to use. ROPE is also sensitive to the scale of measurement. Nonetheless, ROPE can provide much richer information to inform how future studies should be designed.

Extension #2: Cross-level interaction. We can also study the interaction between a macro- and a micro-level variable. Cross-level interaction is useful to study how the effect of context variables manifest differently in individuals with different characteristics. An example of cross-level interaction in our example is whether the relationship between import volume and increase in China coverage manifests differently between public broadcasters and for-profit media.

We can model a cross-level interaction by adding an interaction term between the macro- and micro-level variables, as well as a varying-slope component for the micro-level variable (Heisig & Schaeffer, 2019). With *brms*, it can be done with this:

```
# suppose "public" is a binary variable
import_brms_public <- brm(
  z~offset(log(n))+(1 + public|k)+log(x_import)*public,
  data = outlet_data, family = negbinomial(),
  prior = weaklyinformative_prior, sample_prior = TRUE)
import_brms_public
```

The regression coefficients for the main effect of public broadcasting and the interaction terms have a wide 95% HDI that covers zero. It can be visualized with a conditional effects plot (Figure 5, using the *conditional_effects()* function). The trajectories of public broadcasters and for-profit media outlets are similar, indicating no interaction.

Extension #3: Current study as the informative prior for the next study. Suppose we would like to replicate the study and confirm the same relationship with the 2020 useNews data. This time, we do not need to (and actually, we should not) use the weakly informative prior again. It is because we have *prior* understanding about the

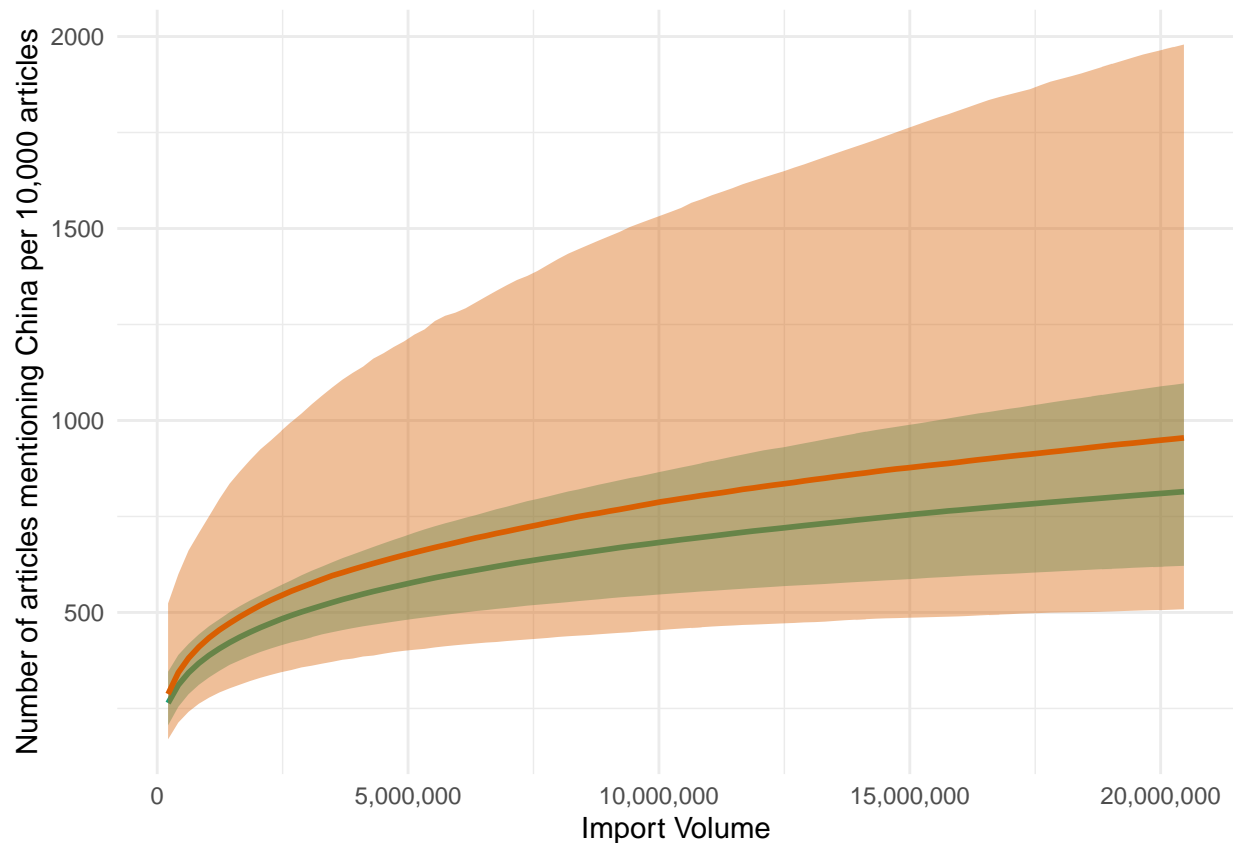


Figure 5. Conditional effects plot showing no interaction among public broadcasting, import volume and China coverage.

outlet-level determinants of China coverage: We know from the 2019 data that γ_{01} should be around 0.25 with a standard deviation of 0.05. This can be entered into our new analysis as the informative prior.

```
informative_prior <- c(prior_string("normal(0.25, 0.05)", class = "b",
                                coef = "logx_import"),
                      prior_string("normal(-6.69, 0.66)",
                                class = "Intercept"))

import_2020 <- brm(z_2020~offset(log(n_2020))+(1|k)+log(x_import),
                  data = outlet_data2020, family = negbinomial(),
                  prior = informative_prior, sample_prior = TRUE)
```


The new analysis showed that γ_{01} is 0.22 with a 95% HDI of (0.15 to 0.29) for the year 2020. This analysis displays two things: 1) selection of prior is not always, as many authors have suggested, as controversial and subjective; and 2) Bayesian analysis also makes a strong case for replication studies, which open science hopes to enable (Dienlin et al., 2020). The use of informative prior in replication studies enables us to update our prior beliefs with the new evidence from the replications. Priors, replicable or not, will get updated.

Extension #4: Temporal changes

Bayesian multilevel regression analysis is also useful to study temporal changes, because observations from the same subject collected from different time points are naturally a cluster. The same strategy for handling the clustering of observations in the multilevel modeling can be used. We model that observations from the same media outlet are nested within the media outlet and then the media outlet is nested within its countries. Suppose we would like to study if there is a systematic increase in z across all outlets, maybe due to the COVID-19 pandemic or the presidential election in the US. It can be analyzed with this three-level model.

```
## Suppose yr is a binary dummy variable indicating the year 2020.
import_long <- brm(z~offset(log(n))+(1|k/i)+yr, data = outlet_long,
  family = negbinomial(),
  prior = weaklyinformative_prior,
  sample_prior = TRUE)
import_long
```

The regression coefficient for the year dummy variable is 0.49 (95% HDI 0.38, 0.60). There is a global increase in China coverage from 2019 to 2020.

Example 2: The Worlds of Journalism study

Example 1 above demonstrated how to conduct a Bayesian multilevel regression analysis. In this example, the same approach is applied to a stochastic dataset from a randomize survey. The starting point of this example is the recent study by Hamada (2021). Using the second wave data from WJS (2012-2016), the study seeks to study this hypothesis (the H2a in the paper): *The greater the level of democracy in a country, the more perceived professional autonomy journalists enjoy.*

It is important to point out that the study is actually a replication. The same hypothesis has been studied in an earlier study by Reich and Hanitzsch (2013), which uses the first wave data from WJS and has a remarkable similar title to that of Hamada (2021). The earlier study has studied this hypothesis (the H4): *Journalists' perceived professional autonomy is positively associated with democratic performance and press freedom, and it is negatively related to political parallelism and state intervention.* The operationalizations of both professional autonomy (based on two questions in WJS) and level of democracy (based on the Economist Intelligence Unit's Index of Democracy) are the same in the two studies.

Interestingly, the analytical approaches are also similar in the two studies. The study by Reich and Hanitzsch (2013) contains multiple regression models on how journalist-level characters predict perceived professional autonomy. However, for country-level predictors such as Index of Democracy, Reich and Hanitzsch (2013) apply an aggregated approach due to “methodological considerations” of “includ[ing] substantive country-level predictors in an OLS regression” (p. 146). Effectively, the analysis boils down to aggregating the perceived professional autonomy of all journalists into mean values according to baskets of Index of Democracy and then comparing those mean values by ANOVA. The subsequent study by Hamada (2021) applies the same aggregated approach by studying the bivariate correlation between the mean perceived professional autonomy of all journalists in a country and the Index of Democracy of a country.

There are two weaknesses of the aggregated approach. First, a massive amount of information at the author-level was discarded and cannot leverage the hierarchical structure of the data. As showed in Reich and Hanitzsch (2013), author-level variables such as year of experience and being in a management position are also predictive of perceived professional autonomy. Taking the aggregated approach assumes that all of these author-level variables are homogeneous. This assumption could create spurious correlation when the distribution of author-level variables is associated with contextual factors, e.g. Index of Democracy. Second, the two hypotheses specified in the two quoted studies refer to the perception at the micro-level (“the more perceived professional autonomy journalists enjoy”, “Journalists’ perceived professional autonomy”), not at the macro-level (e.g. the average level of perceived professional autonomy of journalists in a country). Interpreting the macro-level association wrongly at the micro-level leads to ecological fallacy.

Bayesian analysis of WJS

Let’s assume we were in the shoes of Hamada (2021) and wanted to replicate the study by Reich and Hanitzsch (2013) with the second wave WJS data. Bayesian multilevel regression is probably the only valid approach. First, it can take advantage of the hierarchical structure of the data and correctly estimate the contextual effect of democracy performance at the journalist-level perceived professional autonomy. Second, the incorporating of prior information from Reich and Hanitzsch (2013) makes the intention to replication clearer.

From the model 4 in Reich and Hanitzsch (2013), we know that the standardized beta coefficients for the variables rank and professional experience are .15 and .07 respectively. Unfortunately, the associated standard errors (or standard deviations) were not reported but only asterisks representing statistical significance ($p < 0.001$ for rank, $p < 0.01$ for professional experience). We estimated the maximum standard deviation based on the significance level. For example, the standard deviation of the standardized regression

coefficient ($\sigma_{\beta'_1}$) of rank can be estimated by solving:¹⁰

$$Pr(\frac{\beta'_1}{\sigma_{\beta'_1}}) = \frac{0.001}{2} \sim \mathcal{N}(0, 1) \quad (21)$$

$$\frac{\beta'_1}{\sigma_{\beta'_1}} = 3.29 \quad (22)$$

$$\frac{0.15}{\sigma_{\beta'_1}} = 3.29 \quad (23)$$

$$\sigma_{\beta'_1} = 0.045 \quad (24)$$

or in R

```
beta_d_1 <- 0.15
sig_level <- 0.001
beta_d_1 / abs(qnorm(sig_level / 2))
```

The above information can be used as an informative prior of the current analysis. Unfortunately, the ANOVA result from Reich and Hanitzsch (2013) cannot be used as the prior and we still need to use weakly informative prior. However, Reich and Hanitzsch (2013) suggest that the relationship between Index of Democracy and perceived professional autonomy is in convex “J shape”, with the average perceived professional autonomy of journalists in authoritarian regimes (the lowest end of democracy performance) being higher than those in hybrid regimes (3.92 vs 3.65). We can incorporate this prior information into our model by modeling a cubic relationship between Index of Democracy and perceived professional autonomy. Similar to the general procedure of conducting polynomial regression, we also create a parsimonious model assuming only a

¹⁰ Here we used the z-approximation of Wald test rather than the more typical t-distribution-based Wald test because the degree of freedom (df) is not reported. And t-distribution approximates z-distribution asymptotically when df gets larger.

linear relationship to study whether the cubic regression improves the model fit. In other words, whether the relationship is really in convex “J shape” is studied.¹¹

Another advantage of taking a Bayesian multilevel regression approach rather than the bivariate approach as in Reich and Hanitzsch (2013) is the possibility to adjust for the possible confounding effects of other macro-level variables. For example, GDP per capita (according to World Bank) is moderately correlated with Index of Democracy ($r = 0.53$). The correlation between Index of Democracy and perceived professional autonomy found by Reich and Hanitzsch (2013) could be spurious and the GDP per capita were the actual determinant of perceived professional autonomy. We can adjust for the effect of GDP per capita by entering it also as an independent variable. The data can also provide such variance because there are countries with high GDP per capita but with low democracy (e.g. UAE and Qatar) and vice versa (e.g. Botswana and India).

In order to make the result from this replication comparable to that of Reich and Hanitzsch (2013), we also calculate the standardized regression coefficients, i.e. all variables are mean-centered before entering into the regression model.

The following is the complete code to construct the two Bayesian model. The cubic relationship can be set using the `I()` function.

```
require(tidyverse)

informative_prior <- c(prior_string("normal(0, 1)", class = "b"),
                      prior_string("normal(0, 1)", class = "Intercept"),
                      prior_string("normal(0.14, 0.045)",
                                   class = "b", coef = "EXPRNCE"),
```

¹¹ Actually, the replication by Hamada (2021) has shown that the relationship is not in such convex J shape. But we were assumed to be in the time point before such replication.

```

      prior_string("normal(0.07, 0.027)",
                  class = "b", coef = "RANK"))

std <- function(x){
  (x - mean(x, na.rm = TRUE))/sd(x, na.rm = TRUE)
}

## eiu is Index of Democracy, gdppc is GDP per capita (log)
set.seed(12121)
wjs %>%
  select(ppa, EXPRNCE, RANK, GENDER, UNI_EDU, UNI_EDU, eiu, gdppc, COUNTRY) %>%
  mutate(gdppc = log(gdppc), COUNTRY = as.factor(COUNTRY)) %>%
  mutate_at(vars(ppa:gdppc), std) -> wjs_processed

brm(ppa ~ EXPRNCE + RANK + GENDER + UNI_EDU + eiu + gdppc + (1|COUNTRY),
    data = wjs_processed, cores = 4, iter = 3000, prior = informative_prior,
    sample_prior = TRUE) -> ppa_mod

set.seed(12121)
## iter needs to set larger to ensure convergence
brm(ppa ~ EXPRNCE + RANK + GENDER + UNI_EDU + I(eiu^3) + gdppc + (1|COUNTRY),
    data = wjs_processed, cores = 4, iter = 4000, prior = informative_prior,
    sample_prior = TRUE) -> ppa_qmod

```

By looking at the regression coefficient and the 95% HDI, the standardized regression coefficient for Index of Democracy is 0.24 (95% HDI: 0.13 to 0.34). It is not only statistically significant in the traditional sense but also the strongest predictor among all other predictors (Rank: 0.15, Experience: 0.08, gender: -0.06, university degree: -0.01). It supports Hamada (2021)'s h2a, but our model does that without the problem of

aforementioned ecological fallacy and with adjustment for other potential confounders.

Regarding the possible convex “J shape” relationship suggested by Reich and Hanitzsch (2013), we compare the model fit (R^2) of the cubic model and the parsimonious model and the cubic model provides no improvement in model fit over the parsimonious model (both: 0.194). By looking at the conditional effects plot (Figure 6), it also suggests that the relationship is not convex “J shape”. Journalists in countries scored a lower Index of Democracy have also the lowest perceived professional autonomy.

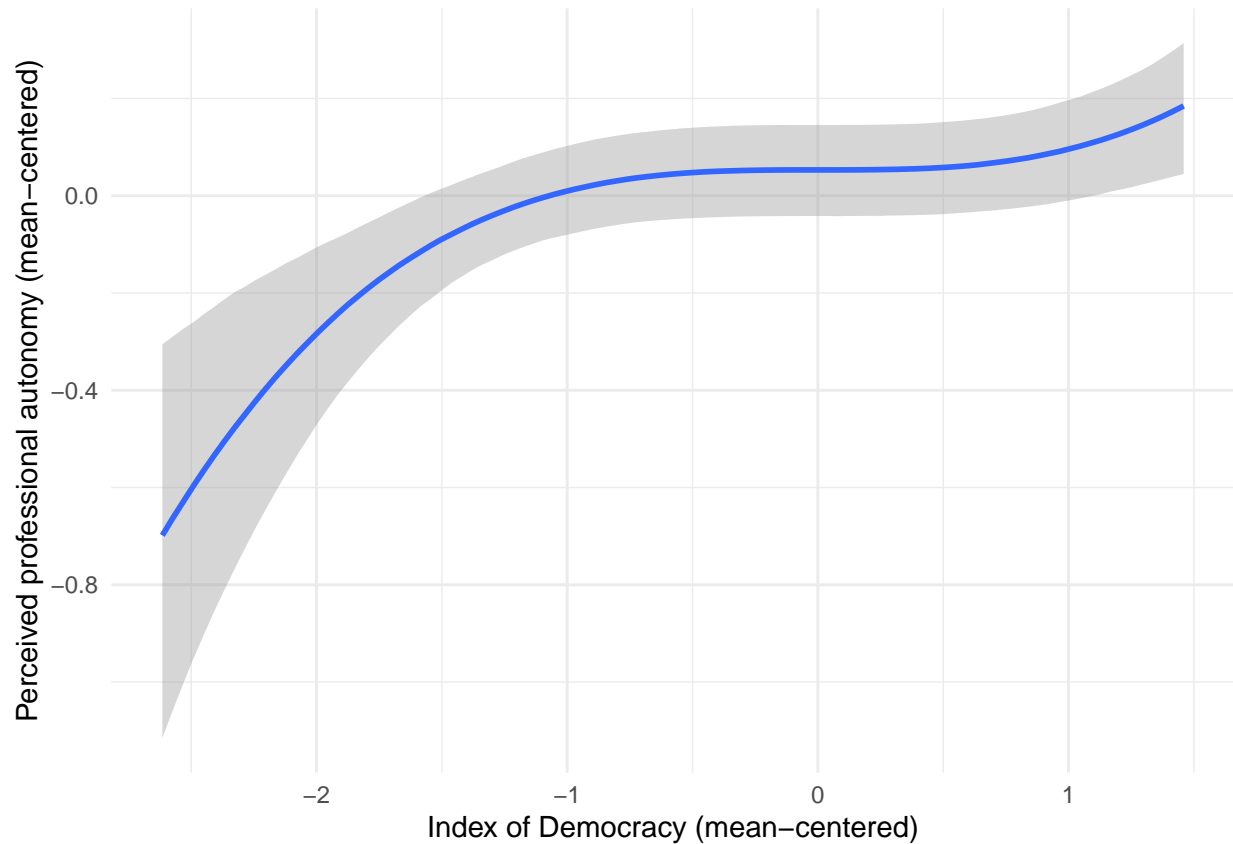


Figure 6. Conditional effects plot showing no convex “J shape” relationship between perceived professional autonomy and Index of Democracy.

Conclusion

In this paper, we argue the case for using Bayesian multilevel regression analysis to analyze data from comparative communication research. We demonstrate using the openly

available useNews and WJS datasets that Bayesian analysis provides valid inference of ecological effects and can be done easily with the R package *brms*. Using a Bayesian approach, however, has also clearly some disadvantages. Researchers have to define priors in order to run a Bayesian regression analysis. As we could show this limitation does not pose a controversial challenge as weakly informative priors can be chosen and in most cases enough data is available and thus the priors have almost no influence on the posterior distribution of the parameters. If we have prior knowledge it allows us to create even better models. Furthermore, it takes more time to get the results of a Bayesian regression analysis. Estimating a Bayesian regression model is a computationally demanding task. It took 4 days on a regular computer to get the results of the Bayesian regression model with 1,525,871 articles (Example 1, Model B1). Still, eventually it worked and typical studies in comparative journalism research (e.g. 6,525 articles in Esser & Umbricht, 2013) can be estimated within a reasonable time frame.¹²

References

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).
<https://doi.org/10.18637/jss.v067.i01>
- Bell, A. V., Richerson, P. J., & McElreath, R. (2009). Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proceedings of the National Academy of Sciences*, 106(42), 17671–17674.
<https://doi.org/10.1073/pnas.0903232106>
- Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent

¹² We benchmarked the same model B1 but with subsets of data of some more “typical” sample sizes. A model with a large sample size of 30,000 can be estimated in less than 17 minutes on a typical 6-core computer. Although it is comparatively more computational intensive than *lme4*, the time needed is still reasonable.

- populations. *Sociological Methodology*, 25, 421. <https://doi.org/10.2307/271073>
- Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* Guilford Press.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455. <https://doi.org/10.1080/10618600.1998.10474787>
- Bryan, M. L., & Jenkins, S. P. (2015). Multilevel modelling of country effects: A cautionary tale. *European Sociological Review*, 32(1), 3–22. <https://doi.org/10.1093/esr/jcv059>
- Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104(3), 396–404. <https://doi.org/10.1037/0033-2909.104.3.396>
- Burggraaff, C., & Trilling, D. (2017). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism*, 21(1), 112–129. <https://doi.org/10.1177/1464884917716699>
- Bürkner, P.-C. (2017). Advanced Bayesian multilevel modeling with the R package brms. *arXiv Preprint arXiv:1705.11123*.
- Chan, C., & Grill, C. (2020). The highs in communication research: Research topics with high supply, high popularity and high prestige in high-impact journals. *Communication Research*. <https://doi.org/10.1177/0093650220944790>
- Chan, C.-h., Wessler, H., Rinke, E. M., Welbers, K., Van Atteveldt, W., & Althaus, S. (2020). How combining terrorism, Muslim, and refugee topics drives emotional tone in online news: A six-country cross-cultural sentiment analysis. *International Journal of Communication*, 14, 26. Retrieved from

<https://ijoc.org/index.php/ijoc/article/view/13247>

Chen, Q. (2012). The impact of ignoring a level of nesting structure in multilevel mixture model. *SAGE Open*, 2(1), 215824401244251.

<https://doi.org/10.1177/2158244012442518>

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health*, 62(8), 752–758. <https://doi.org/10.1136/jech.2007.060798>

de Leeuw, S. E., Azrout, R., Rekker, R. S. B., & Van Spanje, J. H. P. (2020). After all this time? The impact of media and authoritarian history on political news coverage in twelve western countries. *Journal of Communication*, 70(5), 744–767.

<https://doi.org/10.1093/joc/jqaa029>

Deming, W. E., & Stephan, F. F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36(213), 45–49.

<https://doi.org/10.1080/01621459.1941.10502070>

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., ... et al. (2020). An agenda for open science in communication. *Journal of Communication*, 71(1), 1–26. <https://doi.org/10.1093/joc/jqz052>

Durkheim, E. (1897). *Le suicide: étude de sociologie*. F. Alcan.

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, 40(1), 1–5.

Elff, M., Heisig, J. P., Schaeffer, M., & Shikano, S. (2020). Multilevel analysis with few clusters: Improving likelihood-based methods to provide unbiased estimates and accurate inference. *British Journal of Political Science*, 51(1), 412–426.

<https://doi.org/10.1017/s0007123419000097>

- Esser, F., & Umbricht, A. (2013). Competing models of journalism? Political affairs coverage in US, British, German, Swiss, French and Italian newspapers. *Journalism: Theory, Practice & Criticism*, 14(8), 989–1007.
<https://doi.org/10.1177/1464884913482551>
- Esser, F., & Vliegthart, R. (2017). Comparative research methods. *The International Encyclopedia of Communication Research Methods*, 1–22.
<https://doi.org/10.1002/9781118901731.iecrm0035>
- Galtung, J., & Ruge, M. H. (1965). The structure of foreign news. *Journal of Peace Research*, 2(1), 64–90. <https://doi.org/10.1177/002234336500200104>
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), 1360–1383. <https://doi.org/10.1214/08-aoas191>
- Gelman, A., Lee, D., & Guo, J. (2015). Stan. *Journal of Educational and Behavioral Statistics*, 40(5), 530–543. <https://doi.org/10.3102/1076998615606113>
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ...
Modrák, M. (2020). Bayesian workflow. *arXiv Preprint arXiv:2011.01808*.
- Gibbs, B. G., Shafer, K., & Miles, A. (2015). Inferential statistics and the use of administrative data in us educational research. *International Journal of Research & Method in Education*, 40(2), 214–220.
<https://doi.org/10.1080/1743727x.2015.1113249>
- Gnaldi, M., Tomaselli, V., & Forcina, A. (2018). Ecological fallacy and covariates: New insights based on multilevel modelling of individual data. *International Statistical Review*, 86(1), 119–135. <https://doi.org/10.1111/insr.12244>

- Grasland, C. (2019). International news flow theory revisited through a space–time interaction model: Application to a sample of 320,000 international news stories published through RSS flows by 31 daily newspapers in 2015. *International Communication Gazette*, 82(3), 231–259.
<https://doi.org/10.1177/1748048518825091>
- Hamada, B. I. (2021). Determinants of journalists’ autonomy and safety: Evidence from the Worlds of Journalism Study. *Journalism Practice*, 1–21.
<https://doi.org/10.1080/17512786.2021.1871861>
- Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., ... Mascaro, S. (2017). Investigate discuss estimate aggregate for structured expert judgement. *International Journal of Forecasting*, 33(1), 267–279.
<https://doi.org/10.1016/j.ijforecast.2016.02.008>
- Hanusch, F., & Hanitzsch, T. (2017). Comparing journalistic cultures across nations. *Journalism Studies*, 18(5), 525–535. <https://doi.org/10.1080/1461670x.2017.1280229>
- Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review*, 35(2), 258–279. <https://doi.org/10.1093/esr/jcy053>
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5), 1157–1164. <https://doi.org/10.3758/s13423-013-0572-3>
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1), 1593–1623.
- Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

- Keating, D. M., & Tutzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, 43(3), 225–239.
<https://doi.org/10.1080/23808985.2019.1632218>
- Konijn, E. A., van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015). Possible solution to publication bias through bayesian statistics, including proper null hypothesis testing. *Communication Methods and Measures*, 9(4), 280–302.
<https://doi.org/10.1080/19312458.2015.1096332>
- Kovic, M., Rauchfleisch, A., Metag, J., Caspar, C., & Szenogrady, J. (2017). Brute force effects of mass media presence and social media activity on electoral outcome. *Journal of Information Technology & Politics*, 14(4), 348–371.
<https://doi.org/10.1080/19331681.2017.1374228>
- Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280.
<https://doi.org/10.1177/2515245918771304>
- Kundera, M. (1984). *The unbearable lightness of being*.
- Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, 128(7), 912–928.
<https://doi.org/10.1111/oik.05985>
- Makowski, D., Ben-Shachar, M., & Lüdtke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40), 1541. <https://doi.org/10.21105/joss.01541>

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1), 103–123. <https://doi.org/10.3758/s13423-015-0947-8>

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, 2(2), 842–860. <https://doi.org/10.1111/j.1751-9004.2007.00059.x>

Puschmann, C., & Haim, M. (2020). UseNews. Open Science Framework. <https://doi.org/10.17605/OSF.IO/UZCA3>

Reich, Z., & Hanitzsch, T. (2013). Determinants of journalists' professional autonomy: Individual and national level factors matter more than organizational ones. *Mass Communication and Society*, 16(1), 133–156. <https://doi.org/10.1080/15205436.2012.669002>

Rinke, E. M. (2016). The impact of sound-bite journalism on public argument. *Journal of Communication*, 66(4), 625–645. <https://doi.org/10.1111/jcom.12246>

Rinke, E. M., & Schneider, F. M. (2018). Probabilistic misconceptions are pervasive among communication researchers. Open Science Framework. <https://doi.org/10.31235/osf.io/h8zbe>

Robert, C., & Casella, G. (2011). A short history of markov chain monte carlo: Subjective recollections from incomplete data. *Statistical Science*, 26(1), 102–115. <https://doi.org/10.1214/10-sts351>

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t

- tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. <https://doi.org/10.3758/pbr.16.2.225>
- Shikano, S. (2019). Hypothesis testing in the Bayesian framework. *Swiss Political Science Review*, 25(3), 288–299. <https://doi.org/10.1111/spsr.12375>
- Stegmuller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, 57(3), 748–761. <https://doi.org/10.1111/ajps.12001>
- Susser, M. (1994). The logic in ecological: I. The logic of analysis. *American Journal of Public Health*, 84(5), 825–829. <https://doi.org/10.2105/ajph.84.5.825>
- Van Dongen, S. (2006). Prior specification in Bayesian statistics: Three cautionary tales. *Journal of Theoretical Biology*, 242(1), 90–100. <https://doi.org/10.1016/j.jtbi.2006.02.002>
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved \widehat{R} for assessing convergence of mcmc. *Bayesian Analysis*. <https://doi.org/10.1214/20-ba1221>
- Watanabe, K. (2017). Newsmap. *Digital Journalism*, 6(3), 294–309. <https://doi.org/10.1080/21670811.2017.1293487>
- Western, B., & Jackman, S. (1994). Bayesian inference for comparative research. *American Political Science Review*, 88(2), 412–423. <https://doi.org/10.2307/2944713>
- Wilke, J., Heimprecht, C., & Cohen, A. (2012). The geography of foreign news on television. *International Communication Gazette*, 74(4), 301–322. <https://doi.org/10.1177/1748048512439812>
- Wu, H. D. (2000). Systemic determinants of international news coverage: A comparison of

38 countries. *Journal of Communication*, 50(2), 110–130.

<https://doi.org/10.1111/j.1460-2466.2000.tb02844.x>

Zhu, Y., & Fu, K.-w. (2020). Speaking up or staying silent? Examining the influences of censorship and behavioral contagion on opinion (non-) expression in china. *New Media & Society*, 146144482095901. <https://doi.org/10.1177/1461444820959016>