Vương Quân Hoàng, Lã Việt Phương, Trần Trung,
Nguyễn Minh Hoàng, Hồ Mạnh Toàn

BẢN HÒA TẤU
DỮ LIỆU XÃ HỘI

NHÀ XUẤT BẢN KHOA HỌC XÃ HỘI

Vương Quân Hoàng, Lã Việt Phương, Trần Trung,
Nguyễn Minh Hoàng, Hồ Mạnh Toàn

Bản hòa tấu dữ liệu xã hội

NHÀ XUẤT BẢN KHOA HỌC XÃ HỘI
Hà Nội – 2021

Bản hòa tấu dữ liệu xã hội

Nhóm biên soạn

Vương Quân Hoàng

TS. Vương Quân Hoàng (Trung tâm Nghiên cứu Xã hội Liên ngành (ISR), Trường Đại học Phenikaa, Hà Nội) đóng góp hơn 160 nghiên cứu khoa học trên hơn 70 tạp chí quốc tế thuộc nhiều nhà xuất bản lớn như Elsevier, Emerald, MDPI, MIT Press, Nature, Oxford University Press, Palgrave Macmillan, Springer, Taylor & Francis, Wiley, World Scientific.

Kể từ năm 2017 tới nay, ông đã có 13 công bố trên các tạp chí thuộc danh mục xuất bản của Nature Research danh tiếng như *Nature*, *Scientific Data*, *Nature Human Behaviour*. Hiện nay, TS. Vương Quân Hoàng là thành viên ban biên tập các tạp chí thuộc danh mục ISI/Scopus như Editorship: *Humanities & Social Sciences* (Nature Publishing Group); *Scientific Data* (Nature Publishing Group); *Springer Nature Social Sciences* (Springer Nature); *European Science Editing* (European Association of Science Editors);

Lã Việt Phương

Kỹ sư Lã Việt Phương hiện đang tham gia nghiên cứu tại Trung tâm Nghiên cứu Xã hội Liên ngành (ISR), Trường Đại học Phenikaa, Hà Nội. Tính đến nay, anh đã công bố

21 nghiên cứu khoa học.

Trong năm 2019, chương trình bayesvl chạy trong môi trường R do kỹ sư Lã Việt Phương và TS. Vương Quân Hoàng cùng thiết kế và phát triển đã chính thức được chấp nhận công bố trên Comprehensive R Archive Network (CRAN)–thư viện lớn nhất hiện nay dành cho các chương trình trong môi trường R.

Trần Trung

GS.TS. Trần Trung hiện là Giám đốc Học viện Dân tộc thuộc Ủy ban Dân tộc. Ông cũng là Phó Chủ tịch Chi hội Việt Nam, Hiệp hội Các nhà biên tập Khoa học Châu Âu (European Association of Science Editors – EASE), Chủ tịch Hội đồng Biên tập của Tạp chí *Nghiên cứu Dân tộc* (ISSN: 0866-773X), và trưởng nhóm nghiên cứu VSE (Vietnamese Science Editors).

Lĩnh vực nghiên cứu chính của GS.TS. Trần Trung là giáo dục dân tộc, quản lý giáo dục, chính sách công, và phương pháp giảng dạy. Ngoài ra, ông cũng tham gia nghiên cứu liên ngành giữa giáo dục và toán học, khoa học máy tính, kinh tế, công nghệ, và vấn đề phát triển kỹ năng nghiên cứu khoa học.

Nguyễn Minh Hoàng

ThS. Nguyễn Minh Hoàng hoàn thành chương trình thạc sỹ chuyên ngành Khoa học Bền vững tại Trường Đại học Ritsumeikan Châu Á Thái Bình Dương, Nhật Bản. Anh cũng dự kiến sẽ hoàn thành chương trình tiến sỹ tại trường Đại học Ritsumeikan Châu Á Thái Bình Dương vào năm 2022.

Hiện nay, anh đang là nghiên cứu viên tại Trung tâm

Nghiên cứu Xã hội Liên ngành (ISR), thuộc Trường Đại học Phenikaa, Hà Nội. ThS. Nguyễn Minh Hoàng muốn khám phá sâu hơn về các vấn đề liên quan tới nhận thức và tâm lý con người. Anh hy vọng việc giải đáp các vấn đề này sẽ giúp con người đạt được sự bền vững trong nhiều lĩnh vực.

Hồ Mạnh Toàn

ThS. Hồ Mạnh Toàn hoàn thành chương trình thạc sỹ chuyên ngành Kinh tế Phát triển tại Trường Đại học Kinh tế Quốc dân, Hà Nội. Hiện nay, anh đang công tác tại Trung tâm Nghiên cứu Xã hội Liên ngành (ISR), Trường Đại học Phenikaa, Hà Nội. Bên cạnh đó, anh cũng viết các bài truyền thông khoa học cho Hệ thống Truyền thông Khoa học của EASE Việt Nam (`https://sc.sshpa.com/`), cũng như các báo và tạp chí khác tại Việt Nam như *Khoa học và Phát triển*, *VietNamNet*, *Dân Trí*, Tạp chí *Kinh tế và Dự báo*, Tạp chí *Khoa học và Công nghệ Việt Nam*.

Phòng Lab AISDL và Phòng Lab SDAG

Phòng Lab AI for Social Data Lab (AISDL) thuộc công ty TNHH Vương và Cộng sự. Phòng lab được xây dựng với mục tiêu nghiên cứu và phát triển thực nghiệm khoa học xã hội và nhân văn. Thành viên của phòng Lab AISDL gồm có: Đàm Thu Hà, Vương Thu Trang, Vương Hà My, Nguyễn Thanh Nhàn, Nguyễn Thị Linh.

Phòng Lab SDAG (Phân tích dữ liệu trong Khoa học xã hội) là một trong 8 nhóm nghiên cứu mạnh của Trường Đại học Phenikaa. Thành viên của phòng Lab SDAG gồm có: Vương Quân Hoàng, Lã Việt Phương, Hồ Mạnh Tùng, Hồ Mạnh Toàn, Nguyễn Minh Hoàng, Nguyễn Thanh Thanh Huyền, Lê Tâm Trí, Nguyễn Tô Hồng Kông.

Bản hòa tấu dữ liệu xã hội

Mục lục

Bản hòa tấu dữ liệu xã hội

Danh mục hình

Danh mục bảng

Bản hòa tấu dữ liệu xã hội

Từ điển chú giải

Analysis of Variance Phân tích phương sai.. 70

Autocorrelation Dấu hiệu tương quan giữa một chuỗi và bản sao đã làm trễ của chính nó.. 104, 106, 110

Autocorrelation Function (ACF) Hàm tính tự tương quan.. 104, 177

Bayesian Network Lưới Bayesian.. 7

Cohen's d Kích thước hiệu ứng chỉ ra sự khác biệt chuẩn hóa giữa trung bình hai quần thể.. 63, 69

Complete Pooling Quá trình xử lý dữ liệu gộp toàn phần các nhóm dữ liệu trong bộ dữ liệu.. 152, 153

Credibility Mức độ tin cậy.. 50

Effective sample size (ESS) Cỡ mẫu hiệu quả.. 52, 104, 210, 211, 228

Evidence Bằng chứng từ dữ liệu quan sát được, số liệu thu được từ thực nghiệm.. 34

Frequentist Cách gọi phương pháp thống kê cổ điển. Quá trình suy luận của thống kê cổ điển được dựa trên các khái niệm về tần suất của dữ liệu.. 5, 33, 36, 59

Hamiltonian Monte Carlo Thuật toán Hamiltonian Monte Carlo (HMC) là thuật toán lấy mẫu dựa trên nguyên tắc các chuỗi bước được quy định bởi thông tin từ độ dốc bậc nhất.. 180

Highest Density Interval (HDI) Khoảng mật độ cao nhất.. 50, 65

Highest Posterior Density Interval Khoảng mật độ hậu nghiệm cao nhất.. 93

Iterations Bước lặp trong quá trình mô phỏng MCMC.. 48, 52, 110, 175, 210, 213, 228

Lag Độ trễ của chuỗi mô phỏng MCMC.. 104, 105, 110

Likelihood Function Hàm khả năng.. 34, 40, 41, 83

Marginal Likelihood Khả năng biên.. 34

MCMC Thuật toán lấy mẫu bằng cách dựa trên một phân phối xác suất xác định. Phương pháp này sẽ xây dựng một xích Markov với phân phối mong muốn là phân phối cân bằng.. 7, 101, 102, 104, 178

Multilevel Linear Regression Hồi quy tuyến tính đa tầng.. 129, 147, 152

Multiple Linear Regression Hồi quy tuyến tính đa biến.. 129, 147

No Pooling Quá trình xử lý dữ liệu riêng rẽ giữa các nhóm dữ liệu trong bộ dữ liệu.. 152, 153

No-U-Turn Sampler Phần mở rộng của thuật toán Hamiltonian Monte Carlo (HMC). NUTS nhằm loại bỏ bước lựa chọn một số lượng các bước ban đầu của thuật toán HMC.. 180

Ordinary Least Squares (OLS) Phương pháp bình phương nhỏ nhất.. 130

Package Các chương trình máy tính chạy trong môi trường R. Các chương trình này được đóng gói hoàn thiện cùng với hướng dẫn để người dùng có thể sử dụng ngay lập tức.. 33, 51

Partial Pooling Quá trình xử lý dữ liệu gộp riêng các nhóm dữ liệu trong bộ dữ liệu.. 152, 153

Population Quần thể nghiên cứu.. 60

Posterior Probability Xác suất hậu nghiệm.. 33, 40

Prior Probability Xác suất tiền nghiệm.. 34, 40, 42

Probability Density Function Hàm mật độ xác suất.. 41

Probability Mass Function Hàm khối xác suất.. 41

Sample Mẫu nghiên cứu.. 60

Simple Linear Regression Hồi quy tuyến tính đơn giản.. 129, 147

Standard Deviation (SD) Độ lệch chuẩn.. 53

t-test Kiểm định sự khác biệt.. 61

Uncertainty Mức độ bất định.. 3, 34

Unobserved Data Dữ liệu chưa được quan sát.. 33

Varying Intercept Hồi quy tuyến tính hệ số chặn động.. 148, 150

Varying Slope Hồi quy tuyến tính hệ số góc động.. 148

Bản hòa tấu dữ liệu xã hội

Lời giới thiệu

Thưa quý vị độc giả,

Thay mặt nhóm biên soạn, tôi xin chân thành cảm ơn quý vị độc giả đã đón nhận cuốn sách này. Cuốn sách mà các bạn đang cầm trên tay được nhóm biên soạn đúc rút từ chính quá trình làm việc không mệt mỏi trong lĩnh vực nghiên cứu khoa học xã hội và nhân văn (KHXH&NV). Vì vậy, nhóm chúng tôi hy vọng cuốn sách sẽ hỗ trợ đắc lực cho những người muốn dấn thân, cống hiến, và làm việc không mệt mỏi.

Bản hòa tấu dữ liệu xã hội là một trong những nỗ lực sớm nhất trong việc giới thiệu với công chúng một hướng tiếp cận mới trong KHXH&NV: Thống kê Bayesian. Với cách tiếp cận mới, và một phần mềm "cây nhà lá vườn" như bayesvl, chúng tôi hy vọng sẽ giúp những người làm nghiên cứu có thêm công cụ để làm tăng công năng dữ liệu, và làm tăng khả năng tìm tòi, khám phá các ý tưởng độc đáo. Mặc dù bayesvl không phải là chương trình duy nhất về thống kê Bayesian, nhưng với KHXH&NV hiện tại, cách xử lý của bayesvl, mà chúng tôi sẽ trình bày trong suốt 12 chương sách này, là một hướng đi tiên phong.

Chắc hẳn, các bạn cũng có đôi chút tò mò với tên gọi của cuốn sách?

Chúng tôi quyết định đặt tên **Bản hòa tấu dữ liệu xã hội** vì các hình ảnh được vẽ bởi bayesvl làm nhóm biên soạn liên tưởng tới những bản nhạc. Ví dụ như hình Trace

plot (như hình 10.1 ở chương 10), hay hình hệ số Gelman (như hình 10.2, chương 10).

Bên cạnh đó, mỗi công đoạn trong quá trình xử lý thống kê Bayesian với bayesvl như kết nối logic, hay sản xuất hình ảnh đều là một phần của một bản nhạc lớn. Cả quá trình đó, mang tới cho chúng tôi những cảm xúc trầm bổng giống như chơi một bản nhạc. Vì vậy, hy vọng quý vị độc giả cũng sẽ tìm thấy những cảm xúc sâu lắng ấy trong **Bản hòa tấu dữ liệu xã hội** mà chúng tôi giới thiệu với các bạn.

Chúng tôi biên soạn cuốn sách này hướng tới quá trình tự tìm hiểu, sử dụng, và khám phá. Tuy nhiên, có nhiều phần thông tin hiện trạng chưa thực sự đầy đủ, thế nên mục sẽ có nhiều điểm quý vị độc giả cần xem phần 'Tài liệu tham khảo'. Sự thiếu sót này xuất phát từ giới hạn của việc giới thiệu nhiều bài toán bằng một cuốn sách 200 trang. Các bài báo khoa học cung cấp từ 5.000 đến 10.000 chữ để giải quyết một bài toán trọn vẹn. Vì vậy, sự kết hợp giữa cuốn sách và các tài liệu tham khảo là cần thiết cho hành trình tự khám phá cùng bayesvl. Đối với việc sử dụng tài liệu tham khảo, có nhiều tài liệu bằng tiếng Anh nên có thể khiến nhiều độc giả gặp cản trở về ngôn ngữ. Các hạn chế này hy vọng có thể sẽ được khắc phục đầy đủ hơn.

Hiện tại, sau hơn 4 năm nghiên cứu và phát triển, phiên bản 0.8.5 của bayesvl đã được phát hành chính thức trên CRAN tại URL: `https://cran.r-project.org/web/packages/bayesvl/index.html`. Tuy nhiên, phiên bản đầy đủ nhất hiện nay là phiên bản 0.9 trên GitHub, tại địa chỉ: `https://github.com/sshpa/bayesvl`. Bên cạnh việc phát triển phần mềm và tài liệu học tập, trong tương lai, SDAG Lab thuộc Trung tâm Nghiên cứu Xã hội Liên ngành (ISR), Trường Đại học Phenikaa, và phòng Lab AI for Social Data Lab (AISDL) sẽ còn tiếp tục phát triển các khóa học ngắn hạn để hỗ trợ quý vị.

Hy vọng, điều kiện trong tương lai sẽ thuận lợi hơn để chúng tôi có thể được phục vụ quý độc giả trọn vẹn nhất.

Còn bây giờ, xin mời quý độc giả cùng tham gia trải nghiệm, cảm nhận và học tập cùng **Bản hòa tấu dữ liệu xã hội**.

Hà Nội, ngày 10 tháng 7 năm 2021
Thay mặt nhóm tác giả
Vương Quân Hoàng

Bản hòa tấu dữ liệu xã hội

Chương 1

Vài nét sơ lược

1.1 Vài nét sơ lược

Chương trình máy tính **bayesvl** chạy trên môi trường lập trình dành cho ngôn ngữ tính toán thống kê R được chính thức xuất bản trên Comprehensive R Archive Network (CRAN) ngày 24-5-2019, phiên bản v.0.8.5, tại tài liệu số [3]. Các bản cập nhật của chương trình bayesvl được đăng tải miễn phí trên kho mã nguồn mở GitHub. Trên GitHub, hiện tại chương trình đang được đánh số phiên bản v.0.9 [4].

Sau khi vượt qua kiểm tra kỹ thuật của R Core Team, bayesvl được sản xuất thành các phiên bản mã máy, kèm theo tài liệu của R Documentation, cung cấp chính thức tại trang chủ của *CRAN*: `https://cran.r-project.org/web/packages/bayesvl/index.html`. Ngoài Github [4] như đã nói, chương trình có thể dễ dàng tải trực tiếp từ nhiều máy chủ khu vực, đặt tại các trường đại học và trung tâm nghiên cứu liên kết với mạng lưới tài liệu kỹ thuật của R. Vài ví dụ về các nơi có thể tải tài liệu chính thức của bayesvl dưới đây (đã kiểm tra truy cập Ngày 14 tháng 7 năm 2021):

- `https://mran.microsoft.com/package/bayesvl`

- https://cran.csiro.au/web/packages/bayesvl/bayesvl.pdf
- https://rdrr.io/cran/bayesvl/
- https://www.r-pkg.org/pkg/bayesvl
- https://www.crantastic.org/packages/bayesvl
- http://packages.renjin.org/package/org.renjin.cran/bayesvl
- ftp://stat.ethz.ch/CRAN/web/packages/bayesvl/index.html
- https://ftp.uni-sofia.bg/CRAN/web/checks/check_results_bayesvl.html
- http://espejos.ucr.ac.cr/CRAN/web/checks/check_results_bayesvl.html
- http://repo.miserver.it.umich.edu/cran/web/checks/check_results_bayesvl.html
- https://espejito.fder.edu.uy/cran/web/checks/check_results_bayesvl.html

Bên cạnh đó, chương trình cũng được giới thiệu một số nơi như báo *Khoa học & Phát triển* [5], trang web các cơ quan Khoa học - Công nghệ [6], và trường đại học như Trường Đại học Phenikaa [7], hay Trường Đại học Ngoại thương [8].

Chương trình bayesvl phần nào đã trả lời câu hỏi mà nhóm nghiên cứu nêu ra trong nghiên cứu *European Science Editing* vào năm 2019: Làm thế nào để có thể góp phần cổ vũ và phát triển KHXH&NV theo xu hướng hiện đại của thế giới? Làm thế nào để góp phần giải quyết các vấn đề lớn của KHXH&NV hiện nay? Đối với đội ngũ biên

soạn, chúng tôi có niềm tin rằng thúc đẩy khoa học mở, và tiếp cận với thống kê Bayesian có thể sẽ đóng góp những giải pháp hữu ích. Lý do cho nhận định này chúng tôi xin phép được bàn kỹ hơn trong phần sau.

Tuy vậy, trước khi làm quen với thống kê Bayesian và chương trình bayesvl, cuốn sách sẽ bàn sơ lược về các vướng mắc mà ngay cả các nhà nghiên cứu KHXH&NV kỳ cựu vẫn gặp phải khi nhắc tới phương pháp tuy lâu đời nhưng mới mẻ này.

1.2 Những vướng mắc

Sự tồn tại dai dẳng của thói quen nghiên cứu "đếm sao" (Stargazing), *p*-hacking và HARKing [9] là nguyên nhân khiến cho tình trạng nghiên cứu khoa học xã hội không thể tái xác lập trở nên nghiêm trọng [10]. Nhằm giải quyết vấn đề này, Daniel J. Benjamin và cộng sự đã đề xuất thay đổi mức "ý nghĩa thống kê", theo đó *p*-value giảm xuống ngưỡng 0,005 cho những nghiên cứu mới [11]. Đề xuất của họ gần như ngay lập tức gặp phải sự phản đối từ Amrhein & Greenland (2017). Valentin Amrhein và Sander Greenland đưa ra lập luận về mức độ rủi ro cao của việc quá tự tin vào kết quả toán học, việc đánh giá thấp tính không chắc chắn (Uncertainty), và việc loại bỏ các lý luận đơn giản dựa trên "mức ý nghĩa" *p*-value [12].

Thống kê Bayesian có thể là giải pháp cho vấn đề này [13, 14, 15]. Phương pháp này có nhiều điểm phù hợp với khoa học xã hội, ngành khoa học mà mọi quyết định thường được đưa ra dưới sự không chắc chắn [14]. Hơn nữa, phương pháp Bayesian còn có khả năng cập nhật tính hợp lý khi có bằng chứng mới [16]. Đây được coi là sức mạnh cốt lõi của "toán học trên cơ sở thông thường", đặc biệt trong lĩnh vực khoa học xã hội, ngành khoa học có nhiều triết lý không nhất quán và chưa tìm ra giải pháp

cho tình trạng kết quả nghiên cứu không thể tái xác lập. Một thế mạnh khác của phương pháp Bayesian được nhiều nhà khoa học khẳng định là khả năng tinh chỉnh suy luận và sự rõ ràng về sai lệch ước tính [17].

Việc ứng dụng rộng rãi phương pháp thống kê Bayesian là không thể tránh khỏi, đặc biệt khi trở ngại chính của việc áp dụng Bayesian là chi phí tính toán [16] đã được giải quyết nhờ sự phát triển không ngừng của máy tính và các phần mềm mở miễn phí như R, Stan và JAGS [14, 15].

Vấn đề *p*-hacking là vấn đề tồn tại khá lâu trong ngành khoa học xã hội, thậm chí mức α = 0,1 vẫn có thể được chấp nhận trong nhiều ngành con như quản lý hay hành chính công. Johns A. Scales và Roel Snieder (1997) coi sự tồn tại của việc nghiên cứu "đếm sao" trong khoa học xã hội là sự biện minh bằng cách biểu diễn dữ liệu trên cơ sở toán học [16]. Sự tồn tại này làm gia tăng việc ứng dụng nguyên lý entropy cực đại [14, 15]. Nhiều người có khuynh hướng xem những "con số thống kê lộn xộn" là trở ngại cho việc đưa ra "giải pháp thực dụng" trong nghiên cứu [16].

1.2.1 Trở ngại toán học

Ngay cả khi được thuyết phục rằng: "Bạn sẽ không cần sử dụng một chút kiến thức toán nào trong ứng dụng phân tích dữ liệu", thì hầu hết mọi người vẫn coi toán là rào cản tiếp cận với phương pháp Bayesian. Thậm chí những người được đào tạo về thống kê cũng có thể không giải thích chính xác ý nghĩa p = 0,05 rằng 95% giả thiết gốc (null hypothesis) là sai; và giả thiết nghịch là đúng [17]. Dù là với các chuyên gia có kinh nghiệm, các nhà nghiên cứu hiện nay vẫn có thể cần phải tham gia khoá học đào tạo chuyên sâu về kỹ thuật liên quan đến phương pháp Bayesian, trong vòng khoảng 20 đến 50 giờ học (tức khoảng một đến hai học kỳ) để vượt qua nỗi sợ về toán này.

1.2.2 Trở ngại lập trình

Một vấn đề tồn tại đã lâu ở các nhà nghiên cứu khoa học xã hội trong việc áp dụng phương pháp thống kê Bayesian là không có các phần mềm "cắm là chạy" như những phần mềm họ đang quen sử dụng với phương pháp thống kê tần suất hay thống kê cổ điển (Frequentist) [17].

Ngày nay, với sự phổ biến của các ngôn ngữ lập trình mở như R với các phiên bản chạy trên Linux, Windows, MacOS và các công cụ điện toán Bayesian như BUGS, JAGS, Stan không chỉ hoạt động trên R mà còn trên *Mathematica*, *Matlab*; thì việc viết code không nên trở thành vấn đề lớn [14, 15]. Tuy nhiên, các nhà nghiên cứu phải tự mình vượt qua nỗi sợ hãi đó. Nỗi sợ hãi này càng trở nên trầm trọng hơn khi các "phương pháp Fisherian" có vẻ mang lại nhiều lợi ích thiết thực nhờ việc dễ sử dụng, ưu tiên xây dựng mô hình, và cho phép phân công công việc trong việc tìm kiếm giải pháp cho một vấn đề phức tạp [18].

Theo một cách nào đó, việc sử dụng các phần mềm và giải pháp có sẵn trong các vấn đề nghiên cứu cần sử dụng thống kê là một trong những nguyên nhân dẫn đến việc giảm trình độ hiểu biết về máy tính, các kỹ năng lập trình, và quan trọng hơn là giảm khả năng học ngôn ngữ mới để nghiên cứu khoa học.

1.2.3 Vùng an toàn

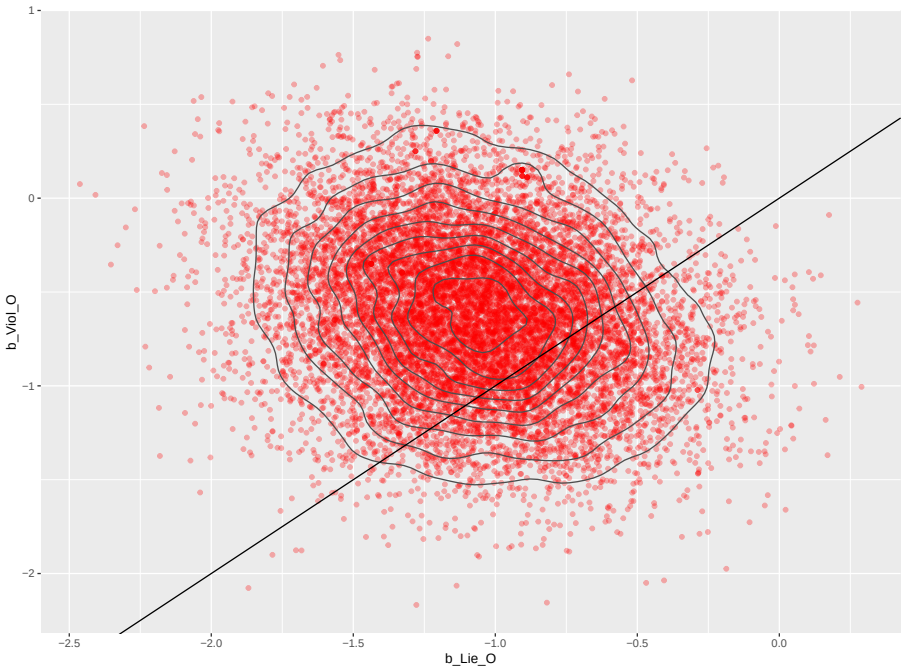
David Malakoff (1999) có lẽ đã đúng khi nhận định lợi ích mà phương pháp thống kê cổ điển (Frequentist) mang lại cho nhà nghiên cứu khoa học xã hội như sau: "Chúng tương đối dễ áp dụng cho các vấn đề thực tế–không giống như phương pháp Bayesian" [17]. Và ở cuối đường hầm, có một cây đũa thần mang tên p-value. Sự kết hợp giữa cây đũa thần p-value và tốc độ tính toán nhanh đã khiến hành vi nghiên cứu "đếm sao" trở thành phương thức chính của

nhiều nhà khoa học; đến mức một số còn tự nói đùa mình là "con khỉ hồi quy".

Vùng an toàn, nơi mà những thói quen của thống kê cổ điển được lặp đi lặp lại mỗi ngày, được bảo vệ mạnh mẽ trước mọi sự xâm nhập từ những khái niệm "ngoài hành tinh" như phương pháp Bayesian. Hiện tại, kể cả khi cuộc khủng hoảng tái xác lập kết quả của ngành tâm lý học [19] đặt ra các vấn đề quan trọng về phương pháp nghiên cứu và xử lý thống kê, việc thay thế hoàn toàn phương pháp thống kê cổ điển bằng Bayesian vẫn chưa được coi là một giải pháp tối ưu. Những nhà nghiên cứu nên chú ý rằng, một nhà khoa học rất thành công như Frank Harrell đã tự chuyển đổi sang phương pháp Bayesian khi ông gặp khó khăn trong việc giải thích vấn đề bằng *p*-value và khoảng tin cậy [20].

1.3 Một số ứng dụng trong công việc

Để giải đáp những vướng mắc được nêu trên đây, nhóm nghiên cứu đã trải qua hành trình tự khám phá và học hỏi cùng với thống kê Bayesian. Cuối cùng, chúng tôi đã được hưởng quả ngọt với các công trình nghiên cứu đã được công bố [21, 22, 1, 23]. Thống kê Bayesian đã tỏ ra hữu dụng cả về xử lý thống kê lẫn cung cấp đồ họa chất lượng (ví dụ, Hình 1.1). Tuy nhiên, để giúp những người mới khác có thể tiếp cận với thống kê Bayesian và triết lý Bayesian mà không gặp những trở ngại do lập trình hay toán học mang lại, nhóm nghiên cứu đã quyết định thiết kế và xây dựng chương trình bayesvl.



Hình 1.1: Một đồ họa sử dụng trong [1]
Nguồn: ©2021 AISDL và SDAG

Trước tiên, chương trình bayesvl không chỉ đơn thuần là một phần mềm thống kê, mà nó còn hỗ trợ người dùng trong việc tiếp cận triết lý Bayesian, cách nghĩ và suy tư về logic và cấu trúc dữ liệu, và cách đánh giá các kết quả thống kê Bayesian. Bên cạnh bayesvl, cũng đã có nhiều chương trình khác có tính năng sư phạm cao như chương trình Rethinking của Richard McElreath đã và đang làm rất tốt cả với phần sách và mã máy tính [14]. Tuy nhiên, tiếp cận tư duy với lưới Bayesian (Bayesian Network), và ước lượng bằng phương pháp MCMC mới thực sự là điểm mạnh của bayesvl, và hiện tại vẫn là ý tưởng riêng của phòng lab AI for Social Data Lab (AISDL).

Một số ví dụ vừa điểm ở trên cho thấy giá trị sử

dụng khá linh hoạt của chương trình bayesvl. Hiện nay, chương trình cũng còn nhiều tiềm năng có thể tiếp tục khai thác, và những điểm cần hoàn thiện cho việc sử dụng được thuận tiện hơn. Trước đây, việc sử dụng chương trình bayesvl còn gặp nhiều hạn chế. Đầu tiên là hạn chế của Bản hướng dẫn sử dụng, bằng cả tiếng Việt [24] và tiếng Anh [25]. Do ban đầu được chuẩn bị để đảm bảo yêu cầu kỹ thuật của R Core Team đối với công tác kiểm tra, đánh giá quy chuẩn chương trình, bản hướng dẫn chỉ tập trung vào các lệnh và khai báo. Các phần hỗ trợ mô hình và lý thuyết tương đối mỏng.

Bên cạnh đó, các tài liệu liên quan tới chương bayesvl cũng thiếu các ví dụ tiêu biểu của các mô hình ứng dụng phân tích thống kê bằng bayesvl. Điều đáng nói là các ứng dụng này phải đảm bảo tiêu chuẩn: đi từ đơn giản đến phức tạp. Nếu tốt hơn nữa, để đảm bảo tin cậy, thì các ứng dụng cụ thể của bayesvl nên được bình duyệt và đã có hiệu chỉnh dựa trên đánh giá của chuyên gia.

Cuối cùng, tài liệu hướng dẫn cũng cần được bổ sung những phần lý giải hoặc so sánh với các mô hình thống kê cổ điển. Mối liên hệ này tỏ ra có ích với những người đã quen với thống kê cổ điển, nhưng hiện còn thiếu những ví dụ toàn diện, đặc biệt là qua dữ liệu và bài toán thực tế. Điều này cho thấy chương trình bayesvl cần có sự bổ trợ của một tài liệu có tính sư phạm tốt hơn (kỹ lưỡng, toàn vẹn và đầy đủ nội dung dẫn chiếu tại chỗ). Chỉ khi có một tài liệu như vậy, khả năng tiếp cận người dùng, hỗ trợ công việc nghiên cứu, giảng dạy, cũng như tìm kiếm hướng để cải thiện tính năng kỹ thuật và thực hành cho người dùng mới có thể được phát huy.

1.4 Hướng giải quyết

Với các vấn đề cụ thể được nêu trên, cuốn sách đặt ra hướng giải quyết các hạn chế như sau:

- Bổ sung các lớp bài toán từ dễ tới khó, sử dụng dữ liệu xã hội có thực, tốt nhất là từ các công bố quốc tế đã qua phản biện kỹ lưỡng. Chuẩn bị các đoạn chương trình cho việc xây dựng các mô hình đó sử dụng bayesvl, và các kết quả kèm theo.
- Sản xuất một giáo trình phục vụ việc đào tạo và sử dụng bayesvl trong công việc nghiên cứu, sao cho 3 ngày có thể đủ cho việc tiếp cận ban đầu và tự đọc tiếp tài liệu liên quan.
- Sử dụng tài liệu đào tạo thử trong phạm vi nhỏ, trực tiếp vào các nghiên cứu thực tế đang triển khai.

Dự án trên bắt đầu ngay từ cuối quý 1 năm 2020. Khi hoàn thành cơ bản thì cũng có lẽ là lúc AISDL sẽ nâng số hiệu phiên bản chương trình lên v.1.0.0. Hy vọng công việc này còn góp phần thúc đẩy tích cực theo hướng chủ động tiếp cận khuynh hướng và nhu cầu phân tích dữ liệu của cộng đồng khoa học quốc tế [26].

Bản hòa tấu dữ liệu xã hội

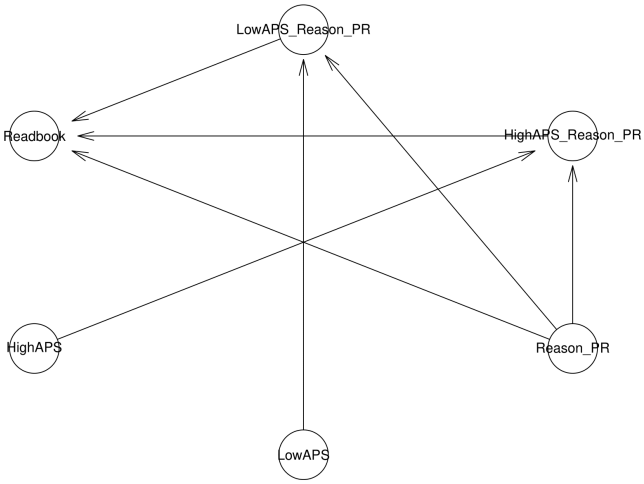
Triển khai và kiểm tra kỹ thuật

Việc triển khai xử lý kỹ thuật ở đây chính là quá trình viết code máy tính trên phần mềm bayesvl. Đối với khoa học hiện đại, việc công khai code máy tính sẽ giúp các nhà nghiên cứu khác có nhiều cơ sở hơn để thực hiện lại nghiên cứu và kiểm tra kết quả [88]. Với mô hình được trình bày ở phần trước, chúng tôi sẽ giới thiệu phần code máy tính như sau:

```
model1<-bayesvl()
model1<-bvl_addNode(model1,"Readbook","binom")
model1<-bvl_addNode(model1,"HighAPS","binom")
model1<-bvl_addNode(model1,"LowAPS","binom")
model1<-bvl_addNode(model1,"Reason_PR","binom")
model1<-bvl_addNode(model1,"HighAPS_Reason_PR","trans")
model1<-bvl_addNode(model1,"LowAPS_Reason_PR","trans")
model1<-bvl_addArc(model1,"HighAPS","HighAPS_Reason_PR",
    "*")
model1<-bvl_addArc(model1,"Reason_PR","HighAPS_Reason_PR
    ","*")
model1<-bvl_addArc(model1,"LowAPS","LowAPS_Reason_PR","*
    ")
model1<-bvl_addArc(model1,"Reason_PR","LowAPS_Reason_PR"
    ,"*")
model1<-bvl_addArc(model1,"Reason_PR","Readbook","slope"
    )
model1<-bvl_addArc(model1,"HighAPS_Reason_PR","Readbook"
    ,"slope")
model1<-bvl_addArc(model1,"LowAPS_Reason_PR","Readbook",
    "slope")
```

Với phần code máy tính được lấy trực tiếp từ R, độc giả của nghiên cứu có thể trực tiếp xây dựng lại mô hình và hiểu sâu hơn về mối quan hệ logic trong việc xây dựng mô hình nghiên cứu. Bên cạnh đó, việc kiểm tra logic và chuẩn mực kỹ thuật của mô hình cũng rất quan trọng. Đồ họa của bayesvl giúp người dùng và cả độc giả có thể kiểm tra các vấn đề này một cách dễ dàng.

Với lệnh bvl_bnPlot, ta sẽ có sơ đồ logic như hình 12.1:



Hình 12.1: Sơ đồ logic của Model
Nguồn: ©2021 AISDL và SDAG

Thông qua biểu thị đồ họa ở hình 12.1, ta có thể dễ dàng kiểm tra các biến và mối quan hệ của chúng. Sau khi kiểm tra tính hợp lý của mô hình, chúng ta sẽ tạo ra code STAN cho mô hình với lệnh bvl_model2stan:

```
functions{
  int numLevels(int[] m) {
    int sorted[num_elements(m)];
    int count = 1;
    sorted = sort_asc(m);
    for (i in 2:num_elements(sorted)) {
      if (sorted[i] != sorted[i-1])
      count = count + 1;
    }
    return(count);
  }
```

```
}
data{
  // Define variables in data
  int<lower=1> Nobs;   // Number of observations (an
   integer)
  int<lower=0,upper=1> Readbook[Nobs];   // outcome
   variable
  int<lower=0,upper=1> HighAPS[Nobs];
  int<lower=0,upper=1> LowAPS[Nobs];
  int<lower=0,upper=1> Reason_PR[Nobs];
}
transformed data{
  // Define transformed data
  vector[Nobs] HighAPS_Reason_PR;
  vector[Nobs] LowAPS_Reason_PR;
  for (i in 1:Nobs) {
    LowAPS_Reason_PR[i] = LowAPS[i]*Reason_PR[i];
  }

  for (i in 1:Nobs) {
    HighAPS_Reason_PR[i] = HighAPS[i]*Reason_PR[i];
  }

}
parameters{
  // Define parameters to estimate
  real a_Readbook;
  real b_Reason_PR_Readbook;
  real b_HighAPS_Reason_PR_Readbook;
  real b_LowAPS_Reason_PR_Readbook;
}
transformed parameters{
  // Transform parameters
  real theta_Readbook[Nobs];
  for (i in 1:Nobs) {
    theta_Readbook[i] = a_Readbook + b_Reason_PR_
   Readbook * Reason_PR[i] + b_HighAPS_Reason_PR_
   Readbook * HighAPS_Reason_PR[i] + b_LowAPS_Reason_PR_
   Readbook * LowAPS_Reason_PR[i];
  }
}
model{
  // Priors
  a_Readbook ~ normal(0,100);
  b_Reason_PR_Readbook ~ normal( 0, 10 );
  b_HighAPS_Reason_PR_Readbook ~ normal( 0, 10 );
```

```
  b_LowAPS_Reason_PR_Readbook ~ normal( 0, 10 );

  // Likelihoods
  Readbook ~ binomial_logit(1, theta_Readbook);
}
generated quantities {
  // simulate data from the posterior
  int<lower=0,upper=1> yrep_Readbook[Nobs];
  // log-likelihood posterior
  vector[Nobs] log_lik_Readbook;
  for (i in 1:num_elements(yrep_Readbook)) {
    yrep_Readbook[i] = binomial_rng(Readbook[i], inv_
   logit(theta_Readbook[i]));
  }
  for (i in 1:Nobs) {
    log_lik_Readbook[i] = bernoulli_logit_lpmf(Readbook[
   i] | theta_Readbook[i]);
  }
}
```

Đây cũng là một phần code quan trọng và nên để vào bài viết nếu có thể. Tuy nhiên, vì phần code STAN này thường rất dài, nên khi trình bày bài, ta có thể để vào phần Tài liệu phụ trợ (Supplementary Materials).

Sau đó, ta tiến hành chạy mô hình với lệnh bvl_modelFit:

```
model1<-bvl_modelFit(model1, data1, warmup = 2000, iter
    = 5000, chains = 4,cores = 4)
summary(model1)
```

Sau khi chạy mô hình thành công, ta sẽ có bảng kết quả như sau xuất hiện trong R.

```
Model Info:
 nodes:     6
 arcs:      7
 scores:    NA
 formula:   Readbook ~ a_Readbook + b_Reason_PR_Readbook * Reason_PR +
    b_HighAPS_Reason_PR_Readbook * HighAPS*Reason_PR + b_LowAPS_Reason
    _PR_Readbook * LowAPS*Reason_PR

Estimates:
Inference for Stan model: e7f9b0a5141d9a85d9c49aff0884df71.
4 chains, each with iter=5000; warmup=2000; thin=1;
post-warmup draws per chain=3000, total post-warmup draws=12000.

                          mean se_mean   sd  2.5%   25%  50%   75%
    97.5% n_eff Rhat
```

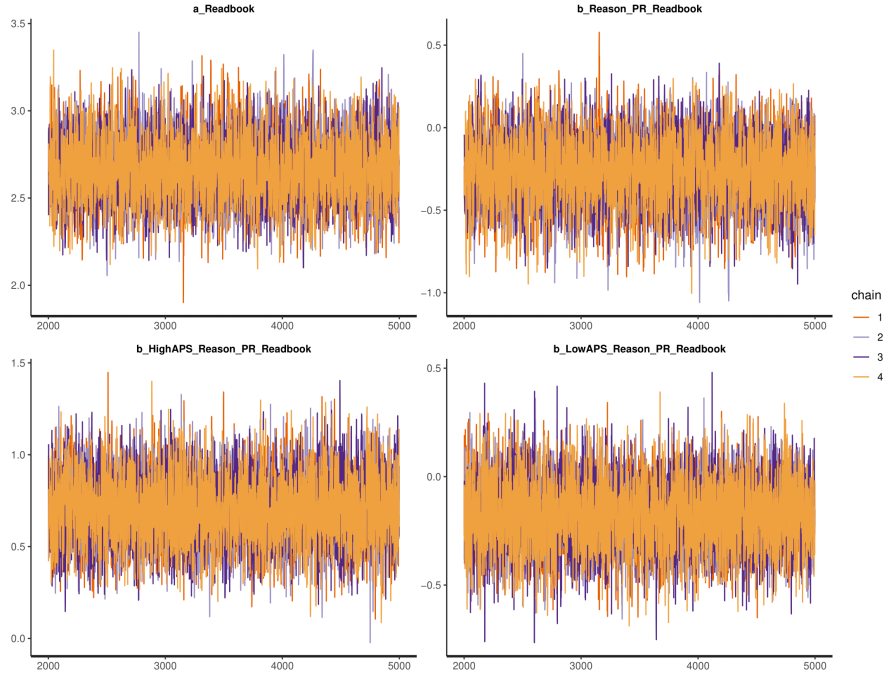
```
a_Readbook                    0.12    0.03 1.78 -3.41 -0.99 0.10  1.18
    3.80  4827    1
b_Reason_PR_Readbook          2.51    0.03 1.96 -1.45  1.27 2.52  3.76
    6.47  4720    1
b_HighAPS_Reason_PR_Readbook 8.31    0.10 6.17 -0.20  3.55 7.13 11.85
    22.65  3632    1
b_LowAPS_Reason_PR_Readbook  7.31    0.10 6.21 -1.34  2.54 6.19 11.00
    22.05  3619    1

Samples were drawn using NUTS(diag_e) at Mon Mar 22 00:56:22 2021. For
    each parameter, n_eff is a crude measure of effective sample size,
     and Rhat is the potential scale reduction factor on split chains
    (at convergence, Rhat=1). elapsed time: 57.9118580818176 secs
```

Bên cạnh phần kết quả chính mà chúng ta sẽ sử dụng trong phần tới. Có một số thông số quan trọng chúng ta cần nêu rõ khi kiểm tra kỹ thuật cho mô hình:

- `chains`
- `iter` (Iterations)
- `warmup`
- `n_eff` (Effective sample size (ESS))
- `Rhat`

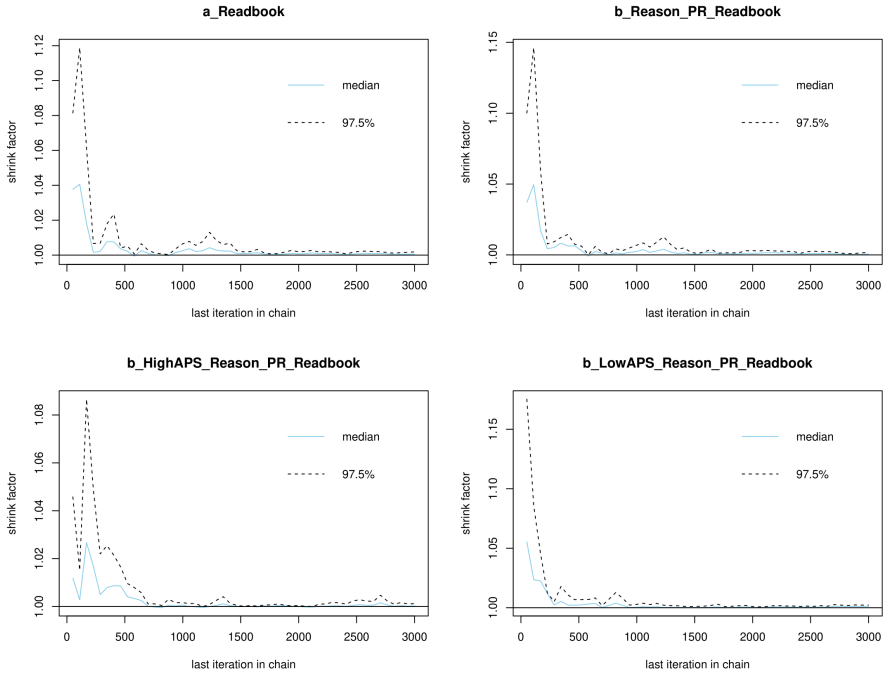
Các thông số này nên được nhắc đến khi trình bày kiểm tra kỹ thuật của mô hình Bayesian. Bên cạnh các thông số, bayesvl và thống kê Bayesian còn có tính chất hình ảnh cao. Các hình ảnh dưới đây là các sản phẩm đồ họa hóa của các cách kiểm tra tiêu chuẩn kỹ thuật của mô hình.

Hình 12.2 kiểm tra hội tụ của các xích Markov. Mỗi xích trong hình đều được chia thành 4 xích thành phần, với 5000 vòng lặp cho mỗi xích (iterations). Về tổng thể, các xích không có chuỗi nào bất thường (divergent chains), tức là bị phân ly và cho thấy dấu hiệu mạnh của hiện tượng tự tương quan (phản tính chất Markov của phân phối). Nếu phân tách một xích thành từng phần và so sánh thì ta sẽ có các phần khá tương đồng nhau về mặt hình ảnh.



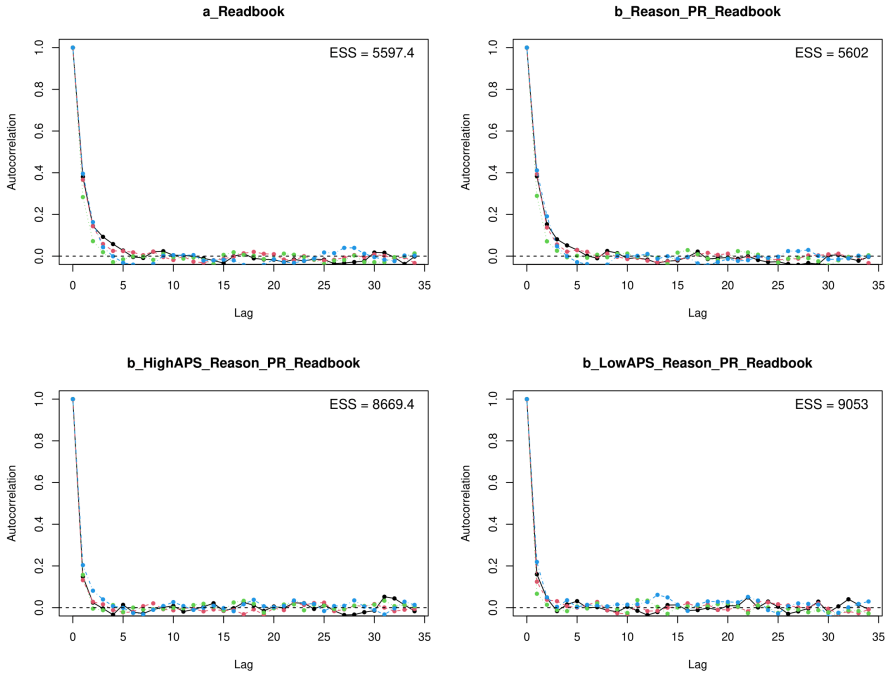
Hình 12.2: Kiểm tra độ hội tụ của các xích Markov
Nguồn: ©2021 AISDL và SDAG

Hình 12.3 cho phép kiểm tra độ hội tụ theo hệ số co Gelman. Hệ số co Gelman ("Gelman shrink factor") hay còn được gọi là hệ số suy giảm quy mô tiềm năng (potential scale reduction factor) thường được sử dụng trong chẩn đoán hội tụ. Chẩn đoán hội tụ này rất cần thiết để đưa ra các kết luận dựa trên phân phối sau, mô tả chính xác mô phỏng tham số và các yếu tố không chắc chắn.



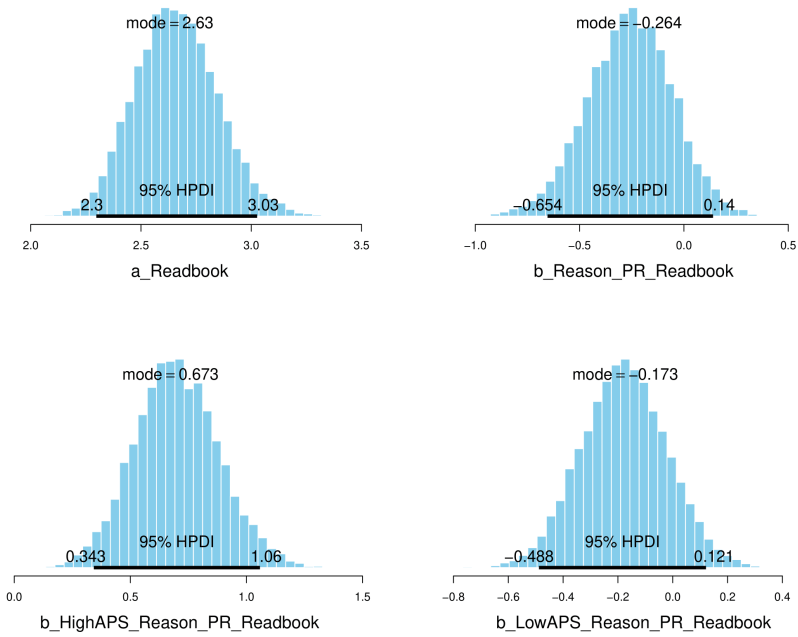
Hình 12.3: Kiểm tra độ hội tụ theo hệ số co Gelman
Nguồn: ©2021 AISDL và SDAG

Hình 12.4 cho phép ta kiểm sự tự tương quan (autocorrelation) của từng hệ số. Thuật toán MCMC sản xuất ra các mẫu tự tương quan (autocorrelation) với nhau chứ không độc lập. Vì vậy, việc bị trộn lẫn (mixing) chậm do tỉ lệ chấp thuận quá cao hoặc thấp có thể dẫn đến các quá trình không đảm bảo tính chất Markov. Việc kiểm tra nhằm đảm bảo sau một số bước hữu hạn, hiện tượng autocorrelation sẽ bị triệt tiêu (về 0).

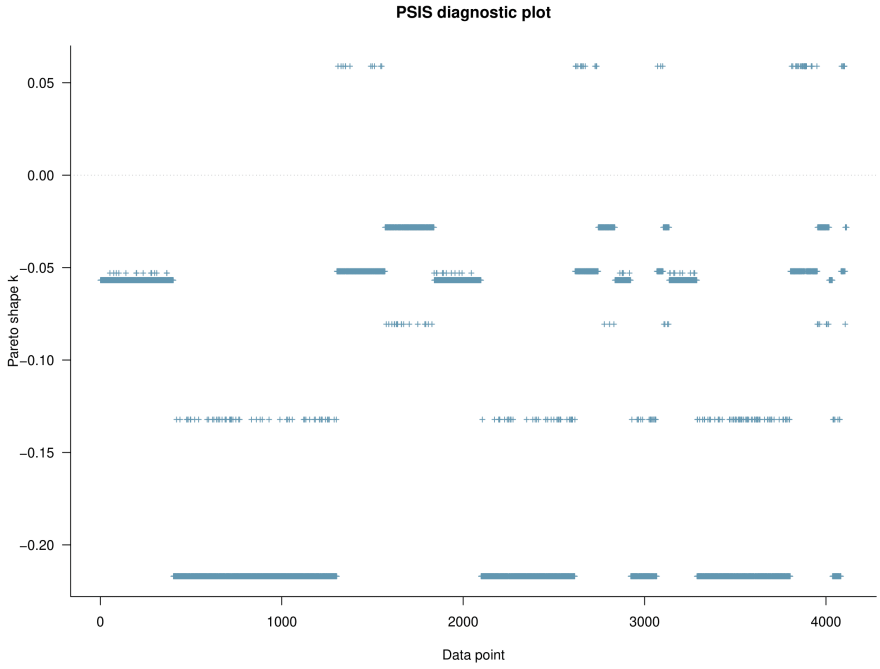


Hình 12.4: Kiểm tra sự tự tương quan của từng hệ số
Nguồn: ©2021 AISDL và SDAG

Hình 12.5 cho thấy mức độ thỏa mãn kỹ thuật của phân phối của các hệ số. Cuối cùng, hình 12.6 cho thấy mức độ phù hợp của mô hình với dữ liệu.



Hình 12.5: Kiểm tra mức độ tin tưởng của phân phối hệ số
Nguồn: ©2021 AISDL và SDAG



Hình 12.6: Kiểm tra mức độ phù hợp của mô hình với dữ liệu
Nguồn: ©2021 AISDL và SDAG

12.1.2 Kết quả và thảo luận

Sau khi chạy mô hình ta sẽ có bảng kết quả bao gồm các thông số kỹ thuật và kết quả chính. Phần kết quả chính đã được bàn đến ở trên. Phần Kết quả này ta sẽ đi sâu vào trình bày kết quả chính. Bảng kết quả mô hình được tóm gọn như sau:

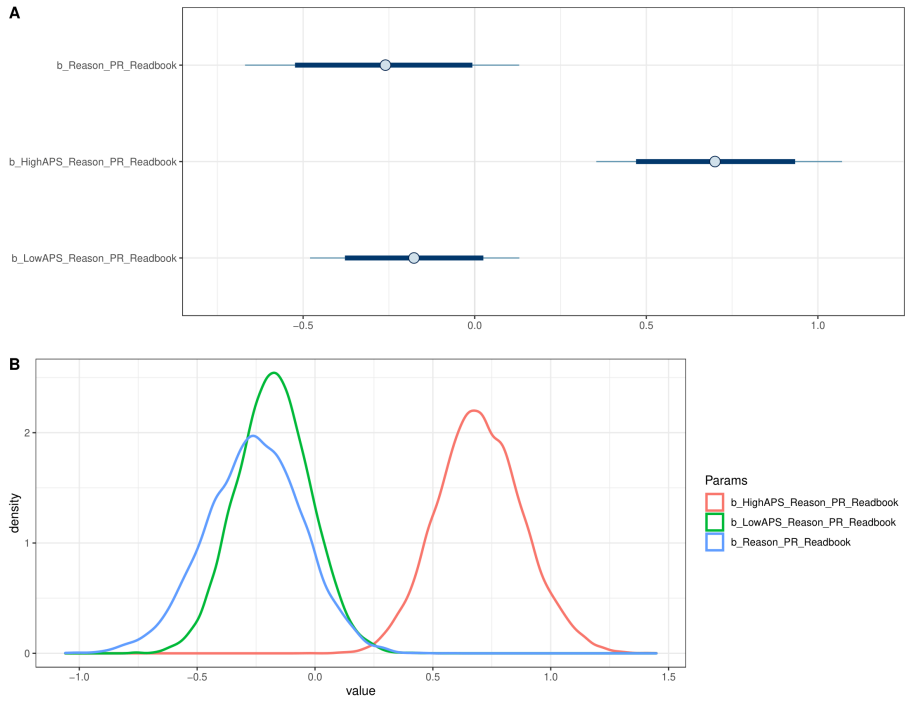
	mean	sd	n_eff	Rhat
a_Readbook	0.12	1.78	4827	1
b_Reason_PR_Readbook	2.51	1.96	4720	1
b_HighAPS_Reason_PR_Readbook	8.31	6.17	3632	1
b_LowAPS_Reason_PR_Readbook	7.31	6.21	3619	1

Bảng 12.2: Ví dụ trình bày một bảng kết quả

Tại bảng 12.2, bốn giá trị chính được trình bày, trong đó 2 giá trị liên quan tới việc kiểm tra kỹ thuật của mô hình:

- Mean: Giá trị Mean của phân phối xác suất hậu mô phỏng.
- Standard Deviation: Độ lệch chuẩn của phân phối xác suất hậu mô phỏng.
- n_eff: Effective sample size.
- Rhat: Các giá trị của Rhat.

Để minh họa cho bảng kết quả chính, hình Intervals hoặc Density được sử dụng. Ví dụ như hình 12.7:



Hình 12.7: Minh họa kết quả trong nghiên cứu [2]
Nguồn: ©2021 AISDL và SDAG

Với các giá trị kết quả ở bảng 12.2 và đồ họa hóa của các kết quả này ở hình 12.7, về cơ bản, ta đã có đầy đủ vật liệu để thảo luận kết quả của một mô hình thống kê Bayesian.

12.2 Bản thảo trước và sau quá trình bình duyệt

Hiện nay, với sự phát triển và tiện dụng của các hệ thống preprints hiện đại [89] thì các bản thảo không chỉ còn dành riêng cho các ban biên tập tạp chí và các nhà bình

duyệt. Một số ý kiến đã chỉ ra các bản thảo ở dạng preprints có những ưu điểm sau (xem kỹ tại [89]):

- Các hệ thống lưu trữ hiện tại đều cung cấp mã số DOI cho các bản preprints, vì vậy các bản này đều có thể được trích dẫn dễ dàng.
- Các bản preprints gốc này giúp lưu trữ ý tưởng gốc của tác giả.
- Cuối cùng, ở góc độ cá nhân, thì các bản thảo là cách thể hiện tinh thần tự do của một người nghiên cứu.

Với tinh thần như trên, việc trình bày một bàn thảo hoàn chỉnh với phương pháp thống kê Bayesian nên tận dụng tối đa sức mạnh về cấu trúc logic và biểu diễn đồ họa. Việc trình bày đẩy đủ và trọn vẹn quá trình phát triển logic, các cơ sở lý thuyết và toán học, kiểm tra kỹ thuật, và kết quả của nghiên cứu sẽ giúp người đọc có thể nắm bắt được trọn vẹn ý tưởng của tác giả. Việc triển khai tốt về mặt kỹ thuật cũng góp phần làm tăng giá trị của nghiên cứu.

Mặt khác, khi bản thảo được trình bày kỹ càng được gửi đi tạp chí, biên tập viên và nhà bình duyệt cũng sẽ khó để có thể từ chối một bản thảo được trình bày kỹ càng. Đồng thời, các nhà bình duyệt cũng có thể dễ dàng đưa ra các góp ý khách quan, tập trung sâu vào nội dung để giúp tác giả khai thác được cái hay của nội dung nghiên cứu.

12.3 Một số nguyên tắc để có bản thảo hay nhất

Tổng kết lại, chúng ta có một số nguyên tắc sau cần chú ý để có được bản thảo hay nhất.

- Dữ liệu mở, code máy tính mở là yếu tố đầu tiên để đảm bảo sự minh bạch của nghiên cứu.
- Cơ sở toán học là yếu tố quan trọng trong việc triển khai mô hình.
- Hình ảnh là sự vượt trội của Bayesian, vì thế cần tận dụng tối đa.
- Bản thảo đầu tiên phải là phiên bản mẫu mực nhất, khai thác tất cả sức mạnh, trình bày đầy đủ, cặn kẽ nhất.

12.4 Khép lại

Đến đây, cuộc hành trình của chúng ta với thống kê Bayesian và phần mềm bayesvl cũng xin được dừng lại. Dựa trên cuộc hành trình của chính đội ngũ biên soạn trong việc áp dụng thống kê Bayesian vào nghiên cứu KHXH&NV tại Việt Nam, cuốn sách khai thác sâu các bài toán thực tiễn, dựa trên các bộ dữ liệu thực tế và phần mềm "cây nhà lá vườn".

Nhóm biên soạn hy vọng người dùng sẽ làm quen với một cách tiếp cận mới, một phương pháp mới, và một triết lý nghiên cứu mới. Khi giới thiệu thống kê Bayesian, điều trước tiên mà chúng tôi hướng tới là một cách suy nghĩ về dữ liệu, và đặt để các vấn đề logic khác so với cách phương pháp thống kê truyền thống. Bên cạnh đó, các xu thế mới trong nghiên cứu khoa học như dữ liệu mở (open data) cũng được giới thiệu và khuyến khích trong cuốn sách này.

Tổng hòa lại, quá trình nghiên cứu khoa học không hề dễ dàng. Một phương pháp mới như Bayesian cũng có thể dễ dàng làm các nhà nghiên cứu khoa học chùn bước. Bản thân nhóm biên soạn cũng hiểu rằng, nội dung cuốn

sách vẫn còn nhiều điểm chưa hoàn mỹ [65]. Mặc dù vậy, khoa học là quá trình tự hoàn thiện liên tục, và mỗi thành tựu khoa học lớn đều được dựa trên những tiến bộ nhỏ hơn. Vì thế, với sự khởi đầu là "*Bản hòa tấu dữ liệu xã hội*" này, cộng đồng KHXH&NV Việt Nam sẽ tiếp tục hướng tới khám phá các hệ giá trị văn hóa, mỹ học và mỹ cảm nói chung của con người [90].

Tài liệu tham khảo

[1] Quan Hoang Vuong et al. On how religions could accidentally incite lies and violence: Folktales as a cultural transmitter. *Palgrave Communications*, 6:82, 2020.

[2] Quan Hoang Vuong et al. Home scholarly culture, book selection reason, and academic performance: Pathways to book reading interest among secondary school students. *OSF Preprints*, 2021, 2021.

[3] Viet Phuong La and Quan Hoang Vuong. bayesvl: Visually learning the graphical structure of Bayesian networks and performing MCMC with 'Stan'. *The Comprehensive R Archive Network (CRAN)*, 2019.

[4] Viet Phuong La and Quan Hoang Vuong. bayesvl package for Bayesian statistical analyses in R. *GitHub*, 2019.

[5] Phương Hạnh Hoàng. Chương trình máy tính bayesvl trong môi trường R: Đóng góp việt cho khoa học thế giới. *Khoa học và Phát triển*, 2019.

[6] Phương Hạnh Hoàng. Chương trình máy tính bayesvl trong môi trường R: Đóng góp việt cho khoa học thế giới. *Trung tâm Thông tin và Thống kê KHCN TP.HCM*, 2019.

[7] Phenikaa University. Phần mềm Việt bayesvl chính thức xuất bản trên cran. *Trường Đại học Phenikaa*, 2019.

[8] YRC. Chương trình máy tính bayesvl trong môi trường R: Đóng góp việt cho khoa học thế giới. *CLB Sinh viên Nghiên cứu Khoa học YRC–FTU*, 2019.

[9] Editorial. Promoting reproducibility with registered reports. *Nature Human Behaviour*, 1:34, 2017.

[10] Manh Tung Ho and Quan Hoang Vuong. The values and challenges of 'openness' in addressing the reproducibility crisis and regaining public trust in social sciences and humanities. *European Science Editing*, 45(1):14–17, 2019.

[11] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2018.

[12] Valentin Amrhein and Sander Greenland. Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, 2(1):4–4, 2018.

[13] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

[14] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. New York: Chapman and Hall/CRC, 2018.

[15] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[16] John A Scales and Roel Snieder. To Bayes or not to Bayes? *Geophysics*, 62(4):1045–1046, 1997.

[17] David Malakoff. Bayes offers a'new'way to make sense of numbers. *Science*, 286(5444):1460–1464, 1999.

[18] Bradley Efron. Why isn't everyone a Bayesian? *The American Statistician*, 40(1):1–5, 1986.

[19] Open Science Framework. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[20] Frank Harrell. My journey from frequentist to Bayesian statistics. *Statistical Thinking*, 2019.

[21] Quan-Hoang Vuong et al. Cultural additivity: behavioural insights from the interaction of Confucianism, Buddhism and Taoism in folktales. *Palgrave Communications*, 4(1):1–15, 2018.

[22] Quan Hoang Vuong et al. Cultural evolution in vietnam's early 20th century: A Bayesian networks analysis of Hanoi Franco-Chinese house designs. *Social Sciences and Humanities Open*, 1(1):100001, 2019.

[23] Manh-Toan Ho et al. An analytical view on STEM education and outcomes: Examples of the social gap and gender disparity in Vietnam. *Children and Youth Services Review*, 119:105650, 2020.

[24] Quan Hoang Vuong and Viet Phuong La. The bayesvl R package. hướng dẫn sử dụng v0.8. *OSF Preprints*, 2019.

[25] Quan Hoang Vuong and Viet Phuong La. The bayesvl R package. user guide v0.8.1. *OSF Preprints*, 2019.

[26] Quan Hoang Vuong. Breaking barriers in publishing demands a proactive attitude. *Nature Human Behaviour*, 3(10):1034–1034, 2019.

[27] Quan Hoang Vuong, Manh Tung Ho, and Viet Phuong La. 'stargazing' and p-hacking behaviours in social sciences: some insights from a developing country. *European Science Editing*, 2019.

[28] Rein Taagepera. *Making social sciences more scientific: The need for predictive models*. OUP Oxford, 2008.

[29] Barbara L Fredrickson and Marcial F Losada. Positive affect and the complex dynamics of human flourishing. *American Psychologist*, 60(7):678, 2005.

[30] Nicholas JL Brown, Alan D Sokal, and Harris L Friedman. The complex dynamics of wishful thinking: The critical positivity ratio. *American Psychologist*, 68(9):801–813, 2013.

[31] Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 2016.

[32] Andrew C. Chang and Phillip Li. Is economics research replicable? sixty published papers from thirteen journals say "often not". *Critical Finance Review*, 7, 2018.

[33] Thomas Herndon, Michael Ash, and Robert Pollin. Does high public debt consistently stifle economic growth? a critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2):257–279, 2014.

[34] Carmen M Reinhart and Kenneth S Rogoff. Growth in a time of debt. *American economic review*, 100(2):573–78, 2010.

[35] Allan Dafoe. Science deserves better: The imperative to share complete replication files. *PS: Political Science and Politics*, 47(1):60–66, 2014.

[36] John Bohannon. Science retracts gay marriage paper without agreement of lead author Lacour. *Science Insider*, 2015.

[37] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016.

[38] Joshua Knobe. Philosophers are doing something different now: Quantitative data. *Cognition*, 135:36–38, 2015.

[39] Joshua D. Greene. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, 167:66–77, 2017.

[40] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021, 2017.

[41] Heather A Piwowar, Roger S Day, and Douglas B Fridsma. Sharing detailed research data is associated with increased citation rate. *PloS ONE*, 2(3):e308, 2007.

[42] Heather A Piwowar and Todd J Vision. Data reuse and the open data citation advantage. *PeerJ*, 1:e175, 2013.

[43] Edwin A. Henneken and Alberto Accomazzi. Linking to data - effect on citation rates in astronomy. *CoRR*, abs/1111.3618, 2011.

[44] Michael J LaCour and Donald P Green. When contact changes minds: An experiment on transmission of support for gay equality. *Science*, 346(6215):1366–1369, 2014.

[45] Jesse Singal. The case of the amazing gay-marriage data: How a graduate student reluctantly uncovered a huge scientific fraud. *New York Magazine*, 2015.

[46] Monya Baker. Stat-checking software stirs up psychology. *Nature*, 540(7631):151, 2016.

[47] Michèle B Nuijten, Chris HJ Hartgerink, Marcel ALM van Assen, Sacha Epskamp, and Jelte M Wicherts. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226, 2016.

[48] Michèle B Nuijten. Preventing statistical errors in scientific journals. *European Science Editing*, 42(1):8–10, 2016.

[49] Marcello Ienca and Effy Vayena. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, 26:463–464, 2020.

[50] Rachel A. McKendry, Geraint Rees, Ingemar J. Cox, Anne Johnson, Michael Edelstein, Andrew Eland, Molly M. Stevens, and David Heymann. Share mobile and social-media data to curb COVID-19. *Nature*, 580:29, 2020.

[51] Esper G Kallas and David H O'Connor. Real-time sharing of Zika virus data in an interconnected world. *JAMA Pediatrics*, 170(7):633–634, 2016.

[52] Anne Holmes, Timothy J. Dallman, Sharif Shabaan, Mary Hanson, and Lesley Allison. Validation of whole-genome sequencing for identification and characterization of shiga toxin-producing escherichia coli to produce standardized data to enable data sharing. *Journal of Clinical Microbiology*, 56(3), 2018.

[53] Editorial. Open for business. *Scientific Data*, 4:170058, 2017.

[54] C. Glenn Begley and Lee M. Ellis. Drug development: Raise standards for preclinical cancer research, 2012.

[55] Quan Hoang Vuong et al. Healthcare consumers' sensitivity to costs: A reflection on behavioural economics from an emerging market. *Palgrave Communications*, 4(1):70, 2018.

[56] Quan Hoang Vuong. Survey data on Vietnamese propensity to attend periodic general health examinations. *Scientific Data*, 4:170142, 2017.

[57] Quan Hoang Vuong et al. An open database of productivity in Vietnam's social sciences and humanities for public use. *Scientific Data*, 5:180188, 2018.

[58] Editorial. Cambridge Analytica controversy must spur researchers to update data ethics. *Nature*, 555:559–560, 2018.

[59] Quan Hoang Vuong. The (ir)rational consideration of the cost of science in transition economies. *Nature Human Behaviour*, 2(1):5, 2018.

[60] Tony Ross-Hellauer. What is open peer review? a systematic review. *F1000Research*, 6:588, 2017.

[61] Tony Ross-Hellauer, Arvid Deppe, and Birgit Schmidt. Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. *PloS ONE*, 12(12):e0189311, 2017.

[62] Gabriele Bammer. What constitutes appropriate peer review for interdisciplinary research? *Palgrave Communications*, 2:16017, 2016.

[63] Declan Butler. Wellcome Trust launches open-access publishing venture. *Nature*, 2016.

[64] Declan Butler. Gates Foundation announces open-access publishing venture. *Nature*, 543(7647):599, 2017.

[65] Quan Hoang Vuong. Reform retractions to make them more transparent. *Nature*, 582(7811):149, 2020.

[66] Stephen J Eglen, Ben Marwick, Yaroslav O Halchenko, Michael Hanke, Shoaib Sufi, Padraig Gleeson, R Angus Silver, Andrew P Davison, Linda Lanyon, Mathew Abrams, et al. Toward standard practices for sharing computer code and programs in neuroscience. *Nature Neuroscience*, 20(6):770–773, 2017.

[67] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press, 2003.

[68] Elie Dolgin. Pubmed Commons closes its doors to comments. *Nature*, 2(2), 2018.

[69] David Mellor. Preregistration and increased transparency will benefit science. *European Science Editing*, 43(4):74–75, 2017.

[70] Andrew Gelman. Objections to Bayesian statistics. *Bayesian Analysis*, 3(3):445–449, 2008.

[71] R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2020.

[72] Nguyen Van Tuan. *Phân tích số liệu và biểu đồ bằng R*. Vienna, Austria: R Foundation for Statistical Computing, 2007.

[73] Emmanuel Paradis. R for beginners, 2002.

[74] Minh Hoang Nguyen et al. Resources for Alice in Suicideland. *Open Science Framework Repository*, 2021.

[75] Manh Toan Ho et al. Health care, medical insurance, and economic destitution: A dataset of 1042 stories. *Data*, 4(2), 2019.

[76] Trung Tran et al. How digital natives learn and thrive in the digital age: Evidence from an emerging economy. *Sustainability*, 12:3819, 2020.

[77] Quan Hoang Vuong et al. On the environment-destructive probabilistic trends: A perceptual and behavioral study on video game players. *Technology in Society*, 65:101530, 2021.

[78] Quan Hoang Vuong et al. Mirror, mirror on the wall: is economics the fairest of them all? An investigation into the social sciences and humanities in Vietnam. *Research Evaluation*, page rvaa036, 2021.

[79] Quan Hoang Vuong et al. Top economics universities and research institutions in Vietnam: evidence from the SSHPA dataset. *Heliyon*, 7(2):e06273, 2021.

[80] Minh Hoang Nguyen et al. Alice in Suicideland: Exploring the suicidal ideation mechanism through the sense of connectedness and help-seeking behaviors. *International*

Journal of Environmental Research and Public Health, 18(7), 2021.

[81] Quan Hoang Vuong et al. Improving Bayesian statistics understanding in the age of big data with the bayesvl R package. *Software Impacts*, 4:100016, 2020.

[82] Quan Hoang Vuong et al. Bayesian analysis for social data: A step-by-step protocol and interpretation. *MethodsX*, 7:100924, 2020.

[83] Quan Hoang Vuong et al. A dataset of Vietnamese junior high school students' reading preferences and habits. *Data*, 4(2), 2019.

[84] Minh Hoang Nguyen et al. A dataset of students' mental health and help-seeking behaviors in a multicultural environment. *Data*, 4(3), 2019.

[85] Quan Hoang Vuong et al. A data collection on secondary school students' STEM performance and reading practices in an emerging country. *Data Intelligence*, 3(2), 2021.

[86] Quan Hoang Vuong and Nancy K. Napier. Acculturation and global mindsponge: An emerging market perspective. *International Journal of Intercultural Relations*, 49:354–367, 2015.

[87] Quan Hoang Vuong. Global mindset as the integration of emerging socio-cultural values through mindsponge processes: A transition economy perspective. In John Kuada, editor, *Global Mindsets: Exploration and Perspectives*, pages 109–126. Routledge, London, 2016.

[88] Quan Hoang Vuong. Open data, open review and open dialogue in making social sciences plausible. *Scientific Data Updates*, 2017.

[89] Quan Hoang Vuong. The rise of preprints and their value in social sciences and humanities. *Science Editing*, 7(1):70–72, 2020.

[90] Quân Hoàng Vương and Ngọc Chiến Bạch. *Bằng chứng cuộc sống: Suy ngẫm về phát triển bền vững Việt Nam*. Nhà xuất bản Chính trị Quốc gia, 2015.

Chỉ mục

Bản hòa tấu dữ liệu xã hội

Bản hòa tấu dữ liệu xã hội

NHÀ XUẤT BẢN KHOA HỌC XÃ HỘI
26 Lý Thường Kiệt - Hoàn Kiếm - Hà Nội
ĐT: 024.39719073 - Fax: 024.39719071
Website: http://nxbkhxh.vass.gov.vn
Email: nxbkhxh@gmail.com

Chi nhánh Nhà xuất bản Khoa học xã hội
57 Sương Nguyệt Ánh - Phường Bến Thành - Quận I - TP. Hồ Chí Minh
ĐT: 028.38394948 - Fax: 028.38394948

Bản hòa tấu dữ liệu xã hội

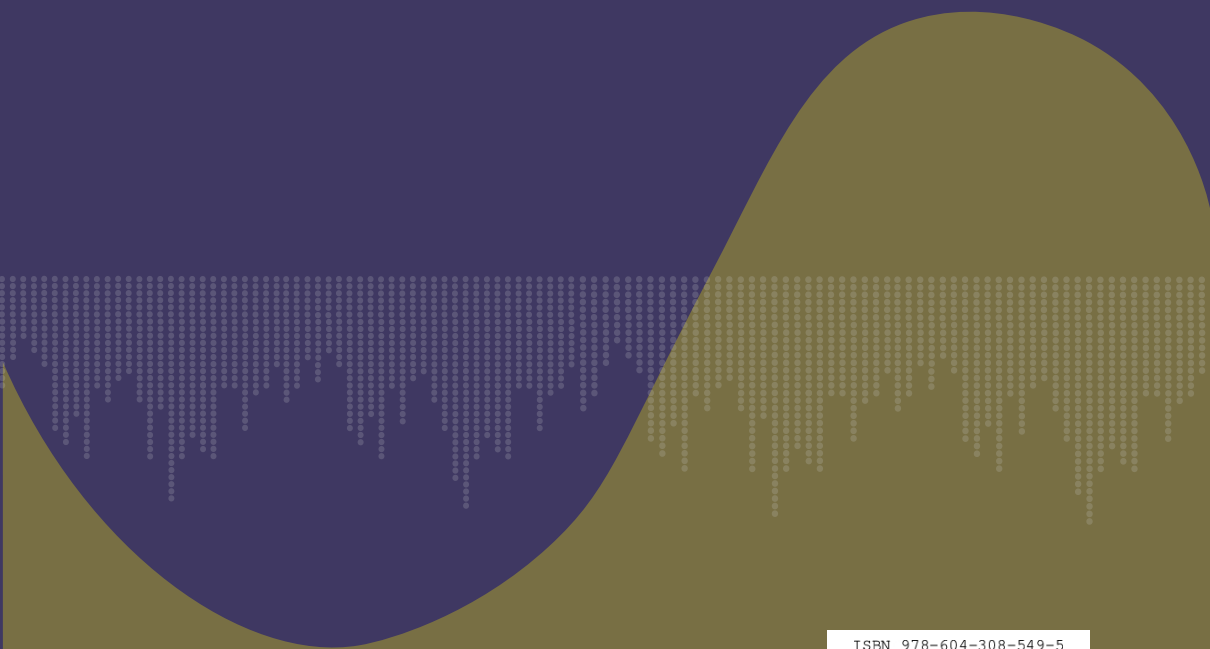
Chịu trách nhiệm xuất bản:
Q. Giám Đốc – Tổng biên tập
PGS.TS. Phạm Minh Phúc

Biên tập nội dung:	Đậu Văn Nam
Chế bản vi tính:	Lã Việt Phương
Sửa bản in:	Hồ Mạnh Toàn
Trình bày bìa:	Nguyễn Quỳnh Anh

In 500 cuốn, khổ 16x24 cm, tại: Công ty Cổ phần thương mại in Hoàng Mai
Địa chỉ: Số 4 Phố Chùa Láng, phường Láng Thượng, quận Đống Đa, thành phố Hà Nội
Số xác nhận đăng ký xuất bản: 2046-2021/CXBIPH/3 - 120/KHXH
Số QĐXB: 132/QĐ-NXB KHXH, ngày 30 tháng 6 năm 2021
ISBN: 978-604-308-549-5.
In xong và nộp lưu chiểu năm 2021.

Chương trình bayesvl được thiết kế với định hướng sư phạm, hỗ trợ các nhà nghiên cứu ngành khoa học xã hội và nhân văn sử dụng mô hình lưới Bayesian, mô phỏng MCMC, hình ảnh hóa các thông số kĩ thuật và kết quả phân tích dữ liệu xã hội. bayesvl được phát hành chính thức trên hệ thống thư viện chuẩn của R là Comprehensive R Archive Network (CRAN) vào tháng 5 năm 2019.

ISBN 978-604-308-549-5



9 786043 085495 >

Giá: 250.000 VNĐ