Is Africa Ready for a New Regional IASC Section? Results and Student Experiences from a Web Scraping Assignment

Joanna Coltrin Adelyn Fleming Cody Hilyard
Jhonatan Medri Rigoberto Tellez Jürgen Symanzik

V Latin American Conference on Statistical Computing (LACSC)
April 21, 2021

Department of Mathematics and Statistics, Utah State University, Logan, UT 84322, USA
E-mails: jcoltrin@aggiemail.usu.edu, adelynflem17@gmail.com, cody.hilyard7@gmail.com, jmedri@aggiemail.usu.edu, tellezrigo12@gmail.com, symanzik@math.usu.edu

Outline

1. Introduction
2. Methods
3. Results
4. Student Feedback
5. Conclusions and References

Project Motivation

- The International Association for Statistical Computing (IASC — `https://iasc-isi.org/`) currently has three regional sections:
  - Europe (IASC–ERS) founded in 1981
  - Asia (IASC–ARS) founded in 1993
  - Latin America (IASC–LARS) founded in 2016
- The IASC was contacted by some of its members to inquire if it would be feasible to establish a new regional section in Africa.
- Provide university students with coursework that has a real-life application.

Research Questions

- Primary Research Question
  - Is Africa ready for a new regional IASC section?
- Secondary Research Questions
  - What is the current activity level of those with a background in statistical computing in Africa?
    - Variables that measure activity level include the number of authors, number of articles, and number of pages.
  - How does the current activity of those in Africa compare to the activity level in Latin America leading up to the creation of LARS?

Project Goal

When the number of members of the IASC is no less than twenty, a newly formed regional section can be approved by the IASC General Assembly.

As of August 2020, there were only 17 African members. While recruiting three or more additional IASC members may be relatively easy, the region must also show it has the potential to conduct typical section activities, such as organizing regional conferences, workshops, and short courses.

The goal of this project is to determine if Africa has enough activity in the field of computational statistics to justify the formation of a new regional section to the IASC.

Data Sources

- To answer these questions, we gathered population information by countries — `http://www.statisticstimes.com/demographics/countries-by-population.php`.
- We then gathered data from two journals in the field of computational statistics:
  - Computational Statistics (COST): `https://www.springer.com/journal/180`
  - Computational Statistics and Data Analysis (CSDA): `https://www.journals.elsevier.com/computational-statistics-and-data-analysis`
- We extracted the year, volume, issue, article title, number of authors, author names, author countries, author order, start page and end page for each article in these journals for the years 2015 to 2020.

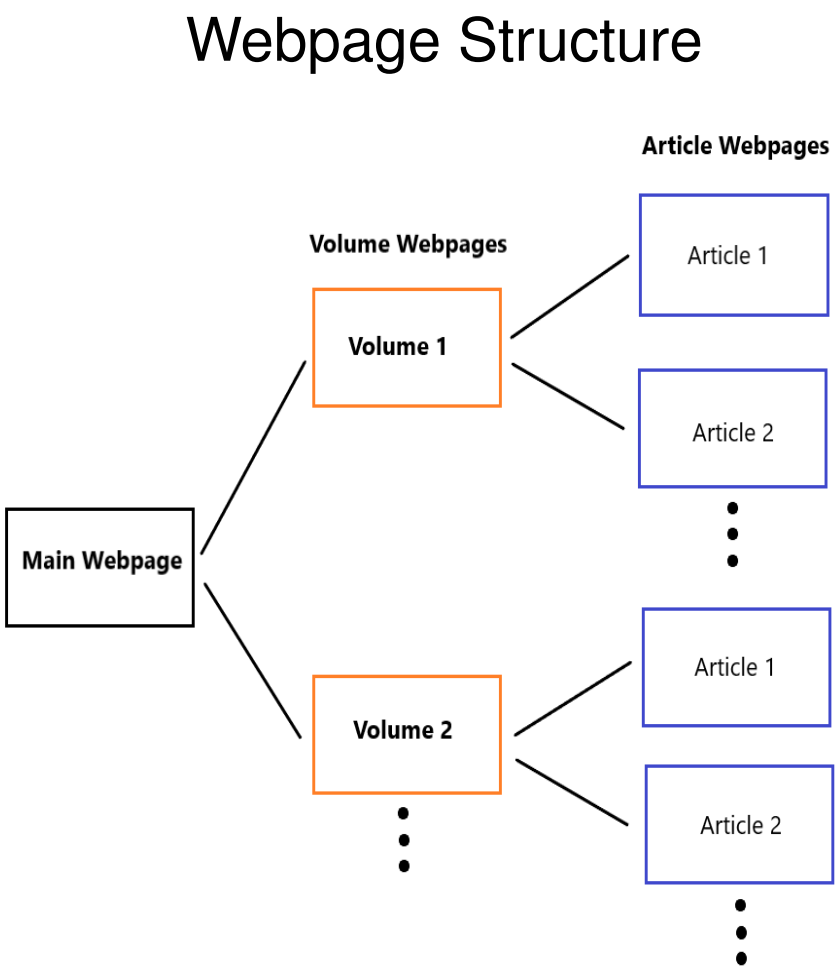
Class Project

- Initially, Utah State University (USU) students registered in the "Data Technologies" course (STAT 5080/6080 — see `https://math.usu.edu/~symanzik/teaching/2019_stat5080/stat5080.html`) in Fall 2019 were assigned this task.
- Five groups of four or five members were created via a stratified random sample (based on previous course work and degree level).
- Each group was asked to collect the data, create meaningful graphics describing these relationships, and determine whether they felt there was enough activity to justify the creation of a regional section in Africa.
- After the completion of the assignment, volunteers from each group created one optimized technique to gather and analyze the data.

Web Scraping

- Web scraping is a modern technique that extracts information from the World Wide Web and compiles it for later use (Hardin et al. 2015; Zhao 2017; Murrell 2009).
- Text patterns, known as regular expressions, allow us to collect and transform unstructured text data to meaningful information (Munzert et al. 2014).
- Both journals (COST and CSDA) had a main webpage with links to journal volumes.
- These volume pages included links to the article webpages which contained the information required.

Depiction of Webpage Structure



Source Code

- HTML coding is the text that creates most webpages.
- This code is accessible via the webpage source code.
- To extract only the relevant information, regular expressions were used after identifying patterns in the text.

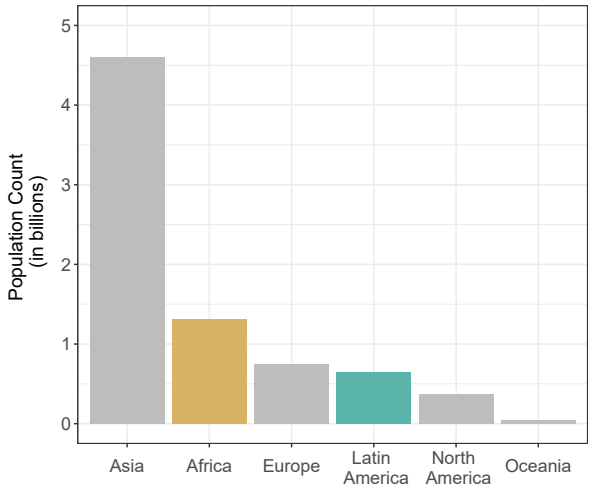
Variable Counts

Variable Counts by Journal and Year for Africa

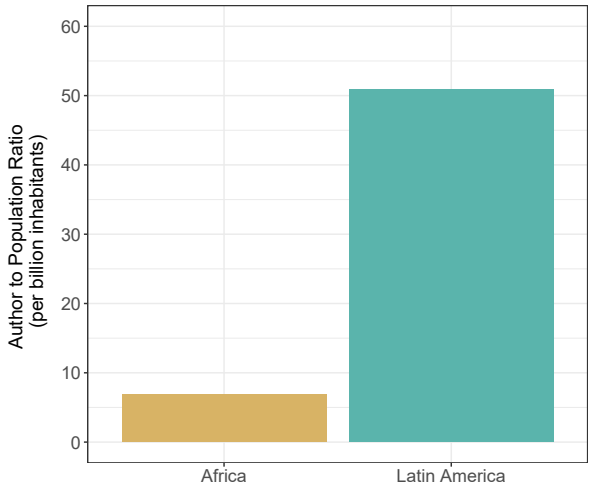
	COST			CSDA		
Year	Articles	Authors	Pages	Articles	Authors	Pages
2015	60	138	1,248	144	381	1,886
2016	75	192	1,619	255	637	3,545
2017	75	184	1,732	188	512	2,593
2018	82	214	1,919	160	446	2,352
2019	82	231	1,867	154	409	2,260
2020	90	231	2,047	168	478	N/A
Total	464	1,190	10,432	1,069	2,863	12,933

World Population Comparison Related to Number of Authors and Pages in 2019

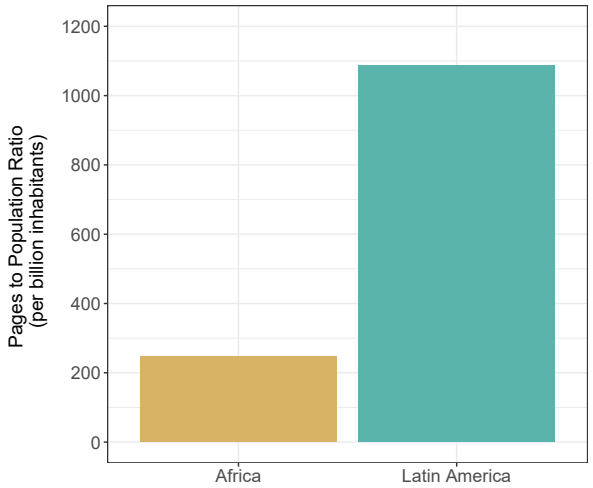
Population Comparison



Author Comparisons

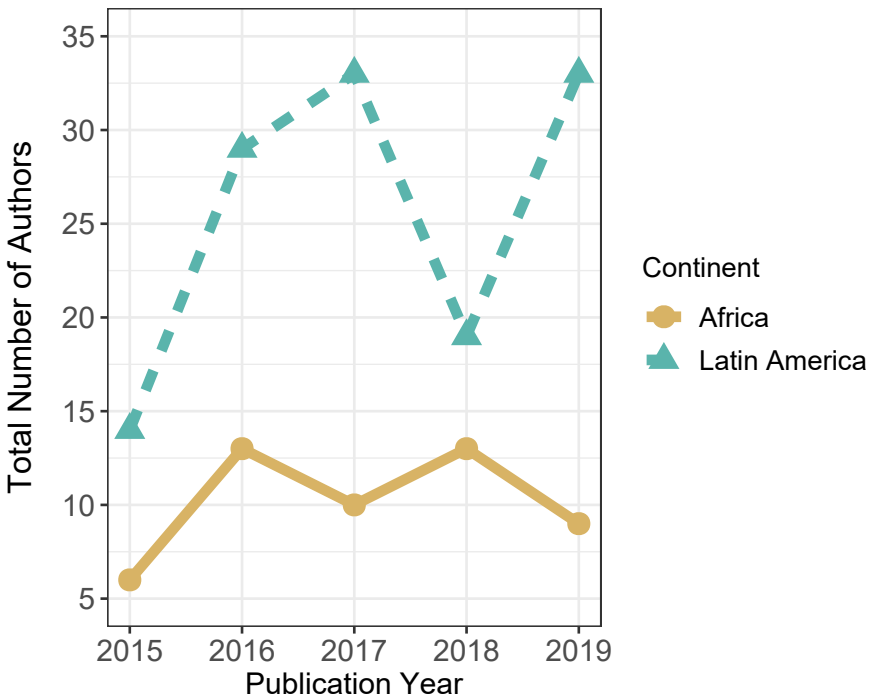


Page Comparisons

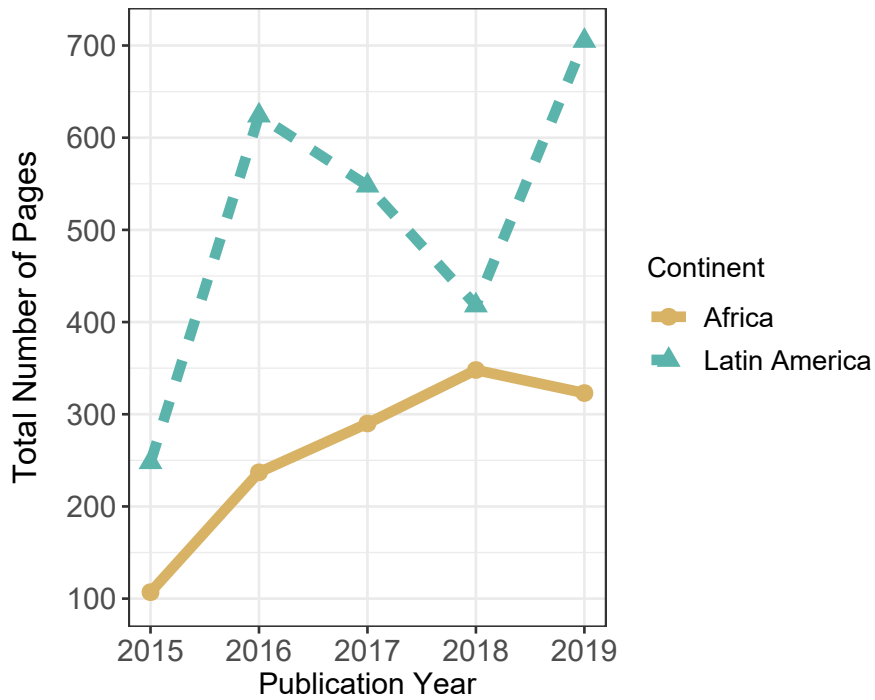


Author and Page Comparisons Over Time

Comparison of Number of Authors

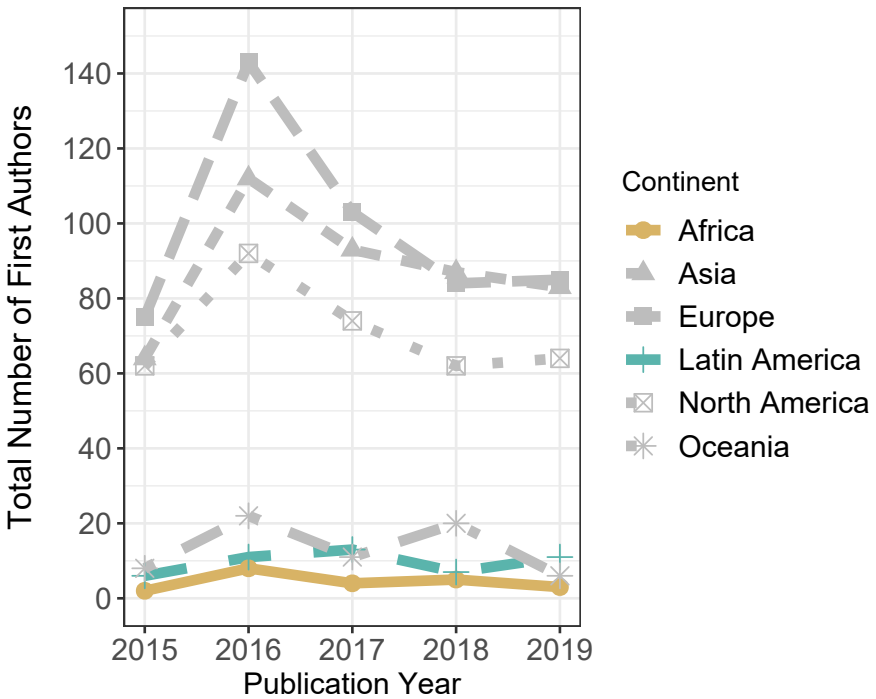


Comparison of Number of Pages

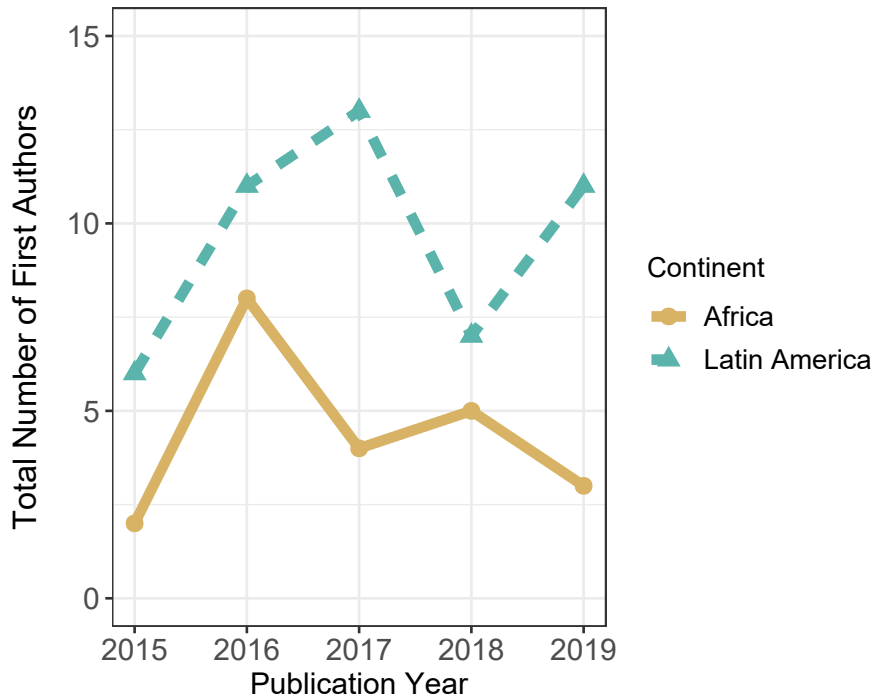


First Author Comparisons

Number of First Author Comparisons



Africa and LARS Comparison



Different Author Weights

- All authors:
  - Count all authors and affiliations with weight 1 (shown here).
  - Count all authors and affiliations with weight $1/m$ for m affiliations.
  - Count all authors and affiliations with weight $1/n$ for n co-authors.
  - Count all authors and affiliations with weight $1/n$ for n co-authors and weight $1/m_i$ for m_i affiliations ($i = 1, \ldots, n$).
- First authors:
  - Count first authors and affiliations with weight 1 (shown here).
  - Count all first authors and affiliations with weight $1/m$ for m affiliations.

General Information

- Tools
  - All groups used R & RStudio.
  - To share data and updated R code, 3 groups used Google Docs; 1 group used Box; 1 group used Dropbox; and 1 group used Excel.
- Time
  - Median estimate: 70 hours (14 hours per student per group).
  - Minimum estimate: 30 to 50 hours.
  - Maximum estimate: 90 hours.

General Information

- Organization
  - Most groups met two or three times as a group and tasks were assigned to individual students. One group met two or three times each week and worked collaboratively on the tasks.
  - To complete the write up, in most groups, one student took the lead while other students contributed shorter parts, based on the work they did. Other students did the proofreading and editing.

Positive Experiences

- Most groups spoke of “extremely relevant”, “very helpful” and something similar to what could be found in a future workplace.
- Many skills were utilized such as data collection (web scraping), data cleaning, data analysis, and visualization.
- The assignment was “open–ended” and had no predetermined “correct” answers. Students learned how to overcome unexpected problems.
- One group stated: “Our group felt that this assignment was way more helpful in the long run than most other types of assignments. We wish we were given more assignments like this in our other college courses as it probably better reflects the types of projects we will be asked to work on in our careers.”

Challenges

- Some undergraduate students felt that this assignment was “a little bit of a stretch.”
- Combining individual R code segments caused some challenges.
- Dealing with time constraints, e.g., when to meet as a group, proved to be difficult for all groups.

Conclusion

- It appears the activity in Africa has increased from 2015 to 2019. Specifically, we see an increase in the number of authors who have published and the number of articles published in computational statistics journals from the African continent. We have also seen an increase in the number of publications from first authors from Africa.
- When we compare the African with the Latin American region, however, we find that these indicators and the rates of increase in these areas are greater in Latin America.
- African authors in 2019 reached a level comparable with authors from Latin America in 2015 (prior to the foundation of IASC–LARS).

References

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D. T., and Ward, M. D. (2015). Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician*, 69(4):343–353.

Munzert, S., Rubba, C., Meißner, P., and Nyhuis, D. (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons, Chichester, UK.

Murrell, P. (2009). *Introduction to Data Technologies*. Chapman and Hall/CRC Press, Boca Raton, FL.

Zhao, B. (2017). Web Scraping. In Schintler, L. A. and McNeely, C. L., editors, *Encyclopedia of Big Data*, pages 1–3. Springer, Cham, Switzerland. https://link.springer.com/referenceworkentry/10.1007%2F978-3-319-32001-4_483-1.

Questions?

or email symanzik@math.usu.edu