EasyChair Preprint

№ 5217

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

Jhonatan Medri, Joanna Coltrin, Adelyn Fleming, Cody Hilyard, Rigoberto Tellez and Juergen Symanzik



March 24, 2021

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

Jhonatan Medri
Utah State University
jmedri@aggiemail.usu.edu

Joanna D. Coltrin
Utah State University
jcoltrin@aggiemail.usu.edu

Adelyn Fleming
Utah State University
adelynflem17@gmail.com

Cody Hilyard
Utah State University
cody.hilyard7@gmail.com

Rigoberto Tellez
Utah State University
tellezrigo12@gmail.com

Jürgen Symanzik *
Utah State University
symanzik@math.usu.edu

March 22, 2021

**Abstract**

In 2019, members of the Executive Committee (EC) of the International Association for Statistical Computing (IASC) were contacted by some of the African IASC members who asked whether it would be feasible to establish a new regional IASC section in Africa. Currently, the IASC has three regional sections: in Europe (IASC–ERS), in Asia (IASC–ARS), and in Latin America (IASC–LARS). To establish a new regional section, there must be a minimum number of IASC members within that geographic region. Moreover, the IASC General Assembly (GA) must approve a new regional section. That approval likely depends on whether the proposed new regional section has the potential to conduct typical section activities, such as organizing regional conferences, workshops, and short courses where most presenters and attendees come from this geographic region. To establish whether it is feasible to add a regional African section, the IASC must know if there is currently enough high-level activity in African countries with respect to computational statistics. To answer this question, we looked at author affiliations of articles published in the *Springer* journal *Computational Statistics* (COST) and the *Elsevier* journal *Computational Statistics & Data Analysis* (CSDA) in the years 2015 to 2019 and used these data as a proxy to compare author productivity for authors with an affiliation in Africa in 2018 & 2019, compared to authors with an affiliation in Latin America in 2015 & 2016 (i.e., immediately before the foundation of IASC–LARS).

In this article, we will look at quantitative results, based on the web scraped author affiliations from COST and CSDA, obtained in an assignment from students

*Corresponding Author

in Utah State University's STAT 5080/0680 "Data Technologies" (DT) course from the Fall 2019 semester.

# 1 Introduction

The International Association for Statistical Computing (IASC) has a "world-wide interest in effective statistical computing and to exchange technical knowledge through international contacts and meetings between statisticians, computing professionals, organizations, institutions, governments and the general public" (see `https://iasc-isi.org/about-iasc2/`). The IASC currently has three regional sections in Europe (IASC–ERS), in Asia (IASC–ARS), and in Latin America (IASC–LARS). The Latin American regional section was founded in 2017. Recently, the IASC was contacted by some of its members with the question whether it would be feasible to establish a new regional section in Africa. 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 (GA). As of August 26, 2020, there were only 17 African members on the non-public internal IASC membership records. While recruiting three or more additional IASC members in Africa may be relatively easy, the GA may approve the formation of a new regional section only if the region shows it has the potential to conduct typical section activities, such as organizing regional conferences, workshops, and short courses where most presenters and attendees come from this geographic region.

In Fall 2019, as part of the "Data Technologies" course at Utah State University (STAT 5080/6080 — see `https://math.usu.edu/~symanzik/teaching/2019_stat5080/stat5080.html`), students were asked to evaluate whether there was enough high-level activity in African countries with respect to computational statistics to justify the creation of a new regional IASC section. Students were assigned via a stratified random sample (based on previous course work and degree level) to one of five groups of four or five members each and asked to gather information from the *Springer* journal *Computational Statistics* (COST — `https://www.springer.com/journal/180`) and the *Elsevier* journal *Computational Statistics & Data Analysis* (CSDA — `https://www.journals.elsevier.com/computational-statistics-and-data-analysis`) to answer this question. The group project focused on the primary question:

Is Africa ready for a new regional IASC section?

In addition, the analyses addressed the following specific questions:

- What is the current activity level of those with a background in statistical computing in Africa?
- How does the current activity of those in Africa compare to the activity level in Latin America leading up to the creation of the Latin American regional section?

This article is structured as follows: In Section 2, we discuss what information was gathered and how it was collected. In Section 3, we present the findings through data visualizations. In Section 4, we discuss the students' perspectives of the group project. In Section 5, we provide our conclusions about the formation of an African regional IASC section and an outlook on future analyses and applications for creative analytics assignments in statistical computing courses. All of our visualizations and analyses are conducted with the `R` statistical computing platform [1].

# 2 Methods

There exist multiple ways to explore the activities of researchers in a geographic region. Traditionally, one might have conducted a phone or mail survey. Alternatively, one could extract information from webpages from university and research institutes. To answer this primary question, we gathered information from two leading journals in the field of computational statistics (COST and CSDA) through web scraping. The information needed from these journals was provided on each journal's web site on a separate webpage for each article published during the period of interest (2015–2019).

Web scraping is a modern technique that extracts information from the World Wide Web and compiles it for later use. One could perform this procedure to acquire specific information from selected web resources. This technique is particularly useful when working with large data sets since the process can be automatized [2, 3, 4].

The extracted information comes in the form of unstructured text with recurring patterns that contain the information we are looking for. These text patterns are also known as regular expressions and allow us to collect and transform unstructured text data to meaningful information [5].

Regarding the class assignment, students were divided into five groups and each group gathered data using a different approach. However, after the completion of the course assignment, students from each of the groups volunteered to assess the web scraping techniques and create one optimized web scraping method. This was done to ensure accuracy of the data to correctly answer the question. So while these methods were employed separately by each of the five student groups initially, the discussion of the web scraping technique used in this article references the collective web scraping process.

## 2.1 Data Gathering

We gathered information from the COST and CSDA journals to answer the primary question. Specifically, we extracted author information from authors located in Africa in the most recent few years and used that to compare to author information from authors located in Latin America in the years proceeding their chapter formation in 2017. We limited the collected data to the years 2015 to 2019.

We initially set out to collect the following information from each research article: journal (COST or CSDA), year, volume, issue, title of article, number of authors for the article, author name, author country, author order, start page of article, and end page of article.

For each of the journals, the web scraping process started on a main webpage that included hyperlinks to specific journal volume webpages, i.e., `https://link.springer.com/journal/180/volumes-and-issues` for COST and `https://www.sciencedirect.com/journal/computational-statistics-and-data-analysis/issues` for CSDA. Each volume page included hyperlinks for individual research articles. The article webpages contained all the information listed above. The structure of these webpages is shown in Figure 1.

The first step in the web scraping process was to gather a list of hyperlinks from the main webpage to access the volume webpages for the years 2015 through 2019. The second step involved taking each hyperlink from the first list and creating a second list of hyperlinks to access the article webpages. Once we had a list of article hyperlinks, we collected the information specified.

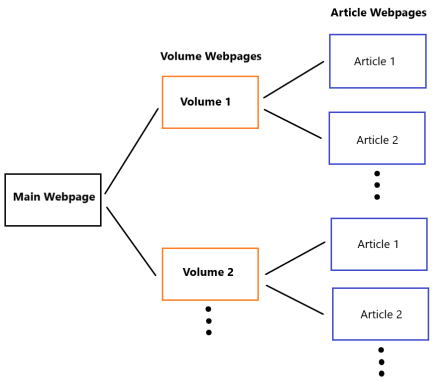



Figure 1: Depiction of the webpage structure for COST and CSDA

Each webpage is created using HTML coding. This HTML code that creates a webpage can be accessed from that webpage's page source. After analyzing the structure of the HTML code from the article webpage page source, we found patterns that helped us identify markers in the HTML code that indicated the article title, author name, author country, and the other information we set out to collect. We used regular expressions to extract only the relevant information from the page source. We compiled this data into a concise table.

We also needed to gather information regarding the number of countries and the total population for 2019. The data was not only gathered for Africa and Latin America, but also for Asia, Europe, North America, Oceania, and South America as well. It was important for us to gather information from all of these continents as a means of comparison when trying to decide whether Africa should move forward with a regional IASC section. The Statistics Times [6] provided a table generated from the World Population Prospects 2019 [7]. The Statistics Times table contains data from all countries from around the world, their 2019 population, and their respective continent. For this project, we used web scraping techniques again to gather this information, grouped by continent and summarized the total number of countries and population count for each individual continent. For the purpose of our group assignments, we assigned all North American countries (except USA and Canada), Central American countries, countries from the Caribbean, and South American countries to Latin America to match the composition of countries that belong to the IASC–LARS regional section.

After web scraping the information from the COST and CSDA journals, we noticed that there were some country names for the authors pulled from the journals that weren't the official country name. Therefore, to make sure we accurately represent each country for the author of the publication, we matched our country names gathered with the naming convention on the CIA World Factbook website (`https://www.cia.gov/the-world-factbook/`).

## 2.2 Data Visualization

To compare the African and Latin American activity levels, we narrowed our focus to the following items: year, number of authors for the article, author country, author

Table 1: Number of articles, number of authors, and number of pages published in COST and CSDA, based on web scraping results

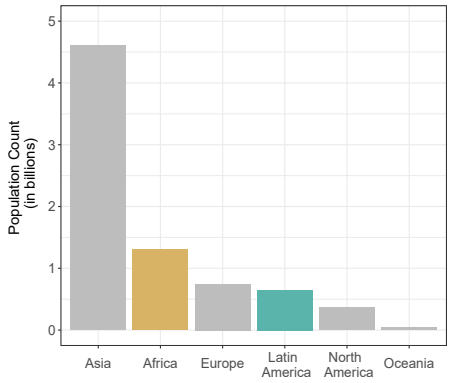
| | COST | | | CSDA | | |
|---|---|---|---|---|---|---|
| **Year** | Articles | Authors | Pages | Articles | Authors | Pages |
| 2015 | 60 | 138 | 1,188 | 144 | 381 | 1,886 |
| 2016 | 75 | 192 | 1,544 | 255 | 637 | 3,545 |
| 2017 | 75 | 184 | 1,657 | 188 | 512 | 2,593 |
| 2018 | 82 | 214 | 1,837 | 160 | 446 | 2,352 |
| 2019 | 82 | 231 | 1,785 | 154 | 409 | 2,260 |
| Total | 374 | 959 | 8,011 | 901 | 2,385 | 12,636 |

order, number of pages per article, and number of pages per author. Looking at these variables specifically helped us to gauge the activity level in Africa. Our graphical exploration includes bar and line charts to depict the interaction between the variables listed above.

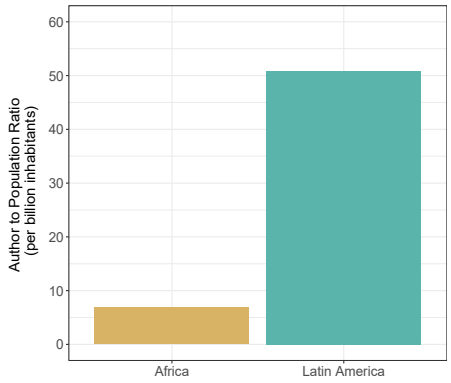
# 3 Results

Our analyses in this section focus on the information gathered from the web scraping process for the years 2015 to 2019. We gathered 374 articles from the COST journal and 901 articles from the CSDA journal. Our procedure effectively accessed 99.84% of the total of 1,275 articles available. Information for the two articles (0.16%) that were not accessed automatically via web scraping because of a system error was manually entered at the end of the process. While we gathered data on all regions, we focused our efforts on comparing the African and Latin American regions in this article. Table 1 shows a summary of those web scraping results. Articles includes all types of articles such as original papers, short notes, and editorials — with the exception of erratums. Authors includes all occurrences of an author, i.e., if someone is an author or co-author of two or three articles in a year, this person is counted two or three times, respectively, in that year. Pages represents the actual number of pages published in a year, but not the final page number of the last issue of that year. This is because of possible blank pages at the end of an article when the journal starts the next article on an odd page number, but not on the available next page. We also refer to first authors in this section. The first author is the author who is listed first in the list of authors and not necessarily the corresponding author. Authors who had listed more than one affiliation were counted for each of these affiliations. This became necessary as there were authors with dual affiliations in Egypt and the United States as well as Ecuador and Belgium for example. Overall, less than 10% of all authors were listed with more than one affiliation.

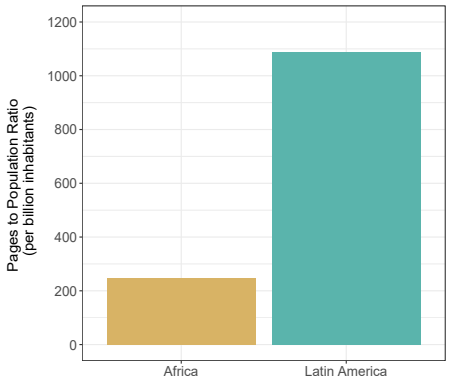
When comparing the African region with the Latin American region, we can make some general observations. Figure 2a shows that the African population of 1.31 billion is about twice as large as the Latin American population of 0.65 billion. While we would expect to see proportions of authors from Africa and Latin America that are comparable to those proportions shown in Figure 2a, Figure 2b shows that the author to population ratio is about nine times greater in Latin America. Figure 2c depicts that the proportion of pages published in Latin America is also much greater than the




(a) World population count (in billions) by continent in 2019




(b) Number of authors to population ratio in Africa and Latin America in 2019



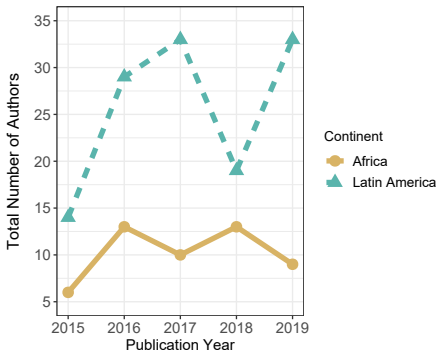

(c) Number of pages to population ratio in Africa and Latin America in 2019

Figure 2: World population in relation to the number of authors and pages published in COST and CSDA

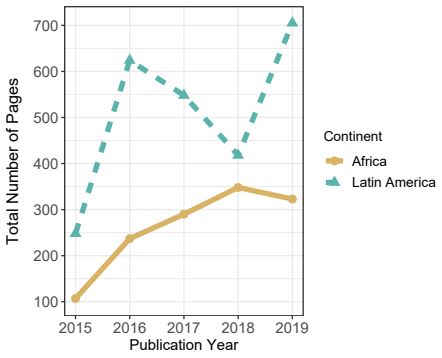
proportion of pages published in Africa when compared to the total population in these two continents.

Figure 3 shows the total number of authors and the total number of pages from Africa and Latin America. In Figure 3a, we can see that both regions have seen some increase in their number of authors who have published in COST and CSDA, but the number of Latin American authors is increasing at a higher rate. Africa has seen an increase of its number of authors who have published in COST and CSDA from six to nine authors while the number of Latin American authors has increased from 14 to 33 over this five-year period.

Figure 3b shows that the total number of pages published by authors from Latin America has been consistently higher than the number of pages published by authors from Africa. It is noteworthy though that the number of article pages published by authors in Africa has been increasing at a greater rate than the number of article pages published by authors in Latin America. Authors in Africa increased their total number of pages from 107 to 323 while authors in Latin America increased their total number of pages from 248 to 705.

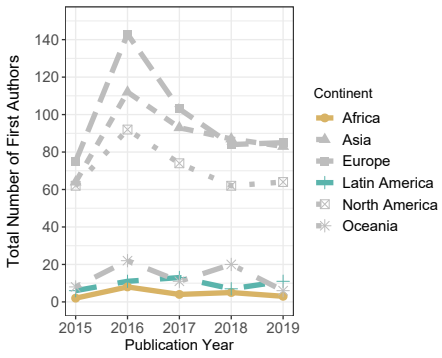



(a) Number of authors who published in COST and CSDA in Africa and Latin America between 2015 and 2019

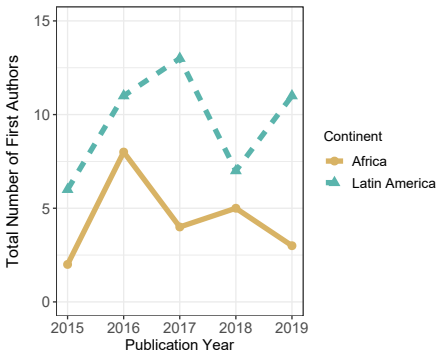



(b) Number of pages published in COST and CSDA in Africa and Latin America between 2015 and 2019

Figure 3: Number of authors and pages published in COST and CSDA between 2015 and 2019




(a) Number of first authors publications in COST and CSDA between 2015 and 2019 by continent




(b) Number of first author publications in COST and CSDA between 2015 and 2019 in Africa and Latin America

Figure 4: Number of first author publications in COST and CSDA between 2015 and 2019

Finally, Figure 4a shows that both African and Latin American regions report the lowest number of first authors publications of all regions. For our web scraping task, we focused on first authors, i.e., authors who are listed first in each research article, and not on the corresponding authors. Taking a look at these statistics is important because it tells us how many research articles were led by a researcher from a certain region. Similar to our previous results, Figure 4b illustrates that Latin American has a consistently higher number of first author publications and that number increases per year at a greater rate as compared to Africa. While Africa has increased the number of first author publications from two to three from 2015 to 2019, Latin American first author publications have grown from six to eleven in the same time frame.

# 4 Students Observations and Feedback

In addition to creating one optimized method for web scraping the data, we were interested in gathering student perspectives of the "Data Technologies" group assignment. For more details on this course and the complete description of this homework assignment, visit `https://math.usu.edu/~symanzik/teaching/2019_stat5080/stat5080.html`. We identified questions that would give relevant and honest answers about the assignment that would benefit any instructors who are interested in assigning similarly complex assignments in their courses. While these answers were gathered from each student in the course, responses were aggregated by group. These responses will be reported for the second part of the conference presentation and the intended journal paper. The questions asked to each group were the following:

- How did you split / organize / recombine the work in your group, e.g., each student working on individual tasks, working together in a physical (or virtual) meeting (if so, how many of these meetings did you have), etc.?
- How did you arrange the writeup for this assignment?
- Which tools did you use, e.g., R, RStudio, box, dropbox, github, Overleaf (an online collaborative LaTeX editor), Google Docs, etc.?
- Did you mostly follow the instructions and order of the question parts in this assignment or did you deviate from it? If so, where? How?
- How relevant was this assignment for your overall DT learning experience?
- What was particularly useful/interesting in this assignment?
- What was not useful/interesting at all in this assignment?
- For those of you who have started a job since the end of the DT course, does any of the material/knowledge from this assignment help in your current job? Which material?
- What, if any, challenges did you face working with a group on an assignment of this magnitude?
- As a group, how many hours did you spend in total on this assignment (i.e., total number of hours for all group members added)?
- Any additional comments?

# 5 Conclusions and Outlook

In conclusion, it appears Africa's current activity 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. While it is possible that the creation of the new regional IASC section created in Latin America in 2017 contributed to the rate of increase, our analysis is

exploratory and doesn't examine causal relationships. It is noteworthy that some of the reported numbers for Africa in 2019 have almost reached or bypassed the same numbers for Latin America in 2015, i.e., prior to the foundation of the IASC–LARS regional section. While the observations from this article do not suggest that an African regional section of the IASC should be created immediately, it suggests that authors from Africa with an interest in computational statistics may reach a capacity in a few years that resembles that in Latin America in 2015 and 2016, i.e., before the foundation of IASC–LARS.

In addition, this article provides some initial insights about how a course project can turn into a research article in which both students and instructors get involved. While we were interested in the web scraped results, the process of obtaining the data through a unique class group project was equally fascinating and should be considered by other instructors teaching computational statistical methods. The feedback portion from the students will be incorporated in the extended journal article following the conference.

# 6 R Tools

Data manipulations and visualizations made use of the `R` packages "boxr" [8], "cowplot" [9], "data.table" [10], "dplyr" [11], "ggplot2" [12], "gridExtra" [13], "httr" [14], "janitor" [15], "kableExtra" [16], "sqldf" [17], "tibble" [18], "XML" [19], and "xtable" [20].

# Acknowledgements

We would like to thank the additional 19 students from the Fall 2019 "Data Technologies" course at Utah State University for their initial contributions to the web scraping efforts and for their answers related to the questions listed in Section 4.

# References

[1] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. `http://www.R-project.org/`.

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

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

[4] P. Murrell. *Introduction to Data Technologies*. Chapman and Hall/CRC Press, Boca Raton, FL, 2009.

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

[6] StatisticsTimes.com. List of Continents by Population, 2020. https://statisticstimes.com/demographics/continents-by-population.php.

[7] United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects 2019, Online Edition. Rev. 1. 2019. https://population.un.org/wpp/.

[8] B. Rocks, I. Lyttle, and N. Day. *boxr: Interface for the 'Box.com API'*, 2019. R package version 0.3.5. https://CRAN.R-project.org/package=boxr.

[9] C. O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2019. R package version 1.0.0. https://CRAN.R-project.org/package=cowplot.

[10] M. Dowle and A. Srinivasan. *data.table: Extension of 'data.frame'*, 2019. R package version 1.12.8. https://CRAN.R-project.org/package=data.table.

[11] H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2020. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr.

[12] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis (2nd Edition)*. Springer-Verlag, New York, NY, 2016. https://ggplot2.tidyverse.org.

[13] B. Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

[14] H. Wickham. *httr: Tools for Working with URLs and HTTP*, 2019. R package version 1.4.1. https://CRAN.R-project.org/package=httr.

[15] S. Firke. *janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2020. R package version 2.0.1. https://CRAN.R-project.org/package=janitor.

[16] H. Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2019. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra.

[17] G. Grothendieck. *sqldf: Manipulate R Data Frames Using SQL*, 2017. R package version 0.4-11. https://CRAN.R-project.org/package=sqldf.

[18] K. Müller and H. Wickham. *tibble: Simple Data Frames*, 2020. R package version 3.0.1. https://CRAN.R-project.org/package=tibble.

[19] D. Temple Lang. *XML: Tools for Parsing and Generating XML Within R and S-Plus*, 2020. R package version 3.99-0.3. https://CRAN.R-project.org/package=XML.

[20] D. B. Dahl, D. Scott, C. Roosen, A. Magnusson, and J. Swinton. *xtable: Export Tables to LaTeX or HTML*, 2019. R package version 1.8-4. https://CRAN.R-project.org/package=xtable.