Computational Statistics
https://doi.org/10.1007/s00180-022-01222-7

ORIGINAL PAPER


# Results and student perspectives on a web-scraping assignment from Utah State University's data technologies course to evaluate the African activity in the statistical computing community

**Adelyn Fleming[1] · Joanna D. Coltrin[1] · Jhonatan Medri[1] · Cody Hilyard[1] · Rigoberto Tellez[1] · Jürgen Symanzik[1]**

Received: 30 August 2021 / Accepted: 21 March 2022


**Abstract**
In 2019, members of the Executive Committee of the International Association for Statistical Computing (IASC) were contacted by members of the IASC from Africa asking whether it would be feasible to establish a new regional IASC section in Africa. The establishment of a new regional section requires several steps that are outlined in the IASC Statutes at https://iasc-isi.org/statutes/. The 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. To establish whether it is feasible to add a regional section in Africa, the IASC must know whether there is currently enough high-level activity within 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) from 2015 to 2020 and used these data as a proxy to compare author productivity for authors with an affiliation in Africa in 2019 and 2020, compared to authors with an affiliation in Latin America in 2015 and 2016. This article looks at quantitative results to the questions above, provides insight on how students from Utah State University's STAT 5080/6080 "Data Technologies" course from the Fall 2019 semester used web scraping techniques in a homework assignment to gather author affiliations from COST and CSDA to answer these questions, and includes the evaluation of student feedback obtained after the end of the course.

**Keywords** International Association for Statistical Computing · IASC · Group project · Data collection · Visualization

✉ Joanna D. Coltrin
joannacoltrin@gmail.com

[1] Department of Mathematics and Statistics, Utah State University, Logan, UT 84322–3900, USA

Published online: 16 April 2022

## 1 Introduction

The International Association for Statistical Computing (IASC) has objectives "to foster 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. These regional sections have been formed in Europe (IASC-ERS), in Asia (IASC-ARS), and in Latin America (IASC-LARS). The Latin American regional section was founded in late 2016 while the Asian and European regional sections were founded in 1993 and 1981, respectively. In 2019, the IASC was contacted by some of its members with the question of whether it would be feasible to establish a new regional section in Africa. When the number of members of the IASC in a particular region is at least twenty, a newly formed regional section can be established (see the IASC Statutes at https://iasc-isi.org/statutes/ for the detailed process). We assumed that a new regional section only gets approved 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. It can be further anticipated that IASC membership in Africa and future activities of its members will have a greater potential to increase after the creation of an African regional section because of formal recruitment activities and events organized by the new section. For this reason, the years 2015 and 2016 have been used as baseline data for IASC-LARS (i.e., prior to the formation of this regional section in late 2016), rather than using the more recent years for a direct comparison with authors from Africa.

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 to one of five groups of four or five members each via a stratified random sample based on previous course work and degree level. The groups were required 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 provide data supporting their position about the IASC's formation of an African regional section. This assignment was presented to the class in the form of a group homework assignment. However, for the purpose of this article, we will be referring to this as a group project because of the complexity involved. This group project was developed to focus on one primary question:

Is Africa ready for a new regional IASC section?

In addition to this primary question, the analysis also 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 Sect. 2, we discuss what information was gathered and how it was collected via web scraping by the students. In Sect. 3, we present the findings through data visualizations. In Sect. 4, we discuss the students' perspectives of the group project. These student perspectives were included because this type of assignment was unique and we felt it would be useful for instructors considering implementing this kind of a project to get a feel for how students felt about the assignment and ways it could have been improved. In Sect. 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 data science, data technologies, and statistical computing courses. Lastly, Appendix A contains the R tools used in our analysis and Appendix B contains the full text of the original assignment provided for our STAT 5080/6080 course.

This article is an extension of the work originally presented in Medri et al. (2021) and includes additional findings for the year 2020, additional weighting methods for the author counts and page counts that are based on Olympic medal table rankings, and an evaluation of the student feedback. All of our visualizations and analyses were conducted with the `R` statistical computing platform (R Core Team 2019).

## 2 Methods

### 2.1 Web scraping

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 our primary question, we gathered information by web scraping. The idea for this web scraping project came from Rundel and Çetinkaya-Rundel (2016) where students were asked to determine the proximity of establishments from a hotel chain and a fast food chain by scraping location information from two websites. 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 across numerous webpages since the process can be automated (Murrell 2009; Hardin et al. 2015; Zhao 2017).

The extracted information comes in the form of unstructured text with recurring patterns that contain the information of interest. These text patterns can be described via regular expressions that allow us to collect and transform unstructured text data into meaningful information (Munzert et al. 2014).

In an effort to answer our primary question, we gathered information from two leading journals in the field of computational statistics, COST and CSDA, through web scraping techniques. The information needed from these journals was provided on each journal's website. These journals list the volumes on individual webpages

that include information for articles contained in that volume. We restricted our data collection about articles to include only those published between 2015 and 2019 for the original assignment and added data from the year 2020 in preparation of this article. It is important to note that the CSDA journal changed to article numbers instead of page numbers in February 2020 (Volume 142) and therefore, our web scraping approach was updated accordingly.

All groups of students were required to use web scraping to collect the necessary information for the assignment. However, each group of students used a different web scraping approach based on the knowledge and skills group members had and how they felt the information could most easily be obtained. 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 for this article. This was done to ensure the accuracy of the data to correctly answer the primary question, rather than to use the results originally obtained from the assignment that differed slightly from group to group. While the web scraping methods were employed separately by each of the five student groups initially, the discussion of the web scraping techniques used for this article references the collective web scraping process of all groups.

### 2.2 Data gathering

Students gathered information from the COST and CSDA journals to answer the primary question: Is Africa ready for a new regional IASC section? Author affiliations for all authors for each article published in COST and CSDA for the years 2015–2019 (and for 2020 in preparation of this article) were obtained via web scraping. Specifically, author affiliations were extracted for those authors located in Africa in the most recent few years. Numbers of authors and page counts were used for comparison with similar information from authors located in Latin America in the years proceeding their regional section formation in late 2016. We targeted the comparison of the African and Latin American geographic regions because of the recent formation of the Latin American section.

We initially set out to collect the following information from each research article for the years 2015–2019 (and later for 2020), excluding erratum articles: respective journal (COST or CSDA), year, volume, issue, title of the article, number of authors for the article, author name(s), author country/countries, author order, start page number of the article, and end page number of the article. We looked at the time span from 2015 to 2020 for a better understanding of the trajectories, in particular for the activities in the African geographic region. The most interesting comparisons between authors in Africa and Latin America could be based on a two-year timeframe that compares only the years 2015 and 2016 for LARS (i.e., prior to the formation of this regional section in late 2016) and for 2018 and 2019 (at the time of the course project) and for 2019 and 2020 (in preparation of this article). We assumed that there might be a formal request to establish a new regional IASC section in Africa in 2020 or 2021, respectively, so that the most recent data for the African geographic region could be used for comparison with the historic data from Latin America. The reader should also be reminded that the web scraping was done as part of a course project where

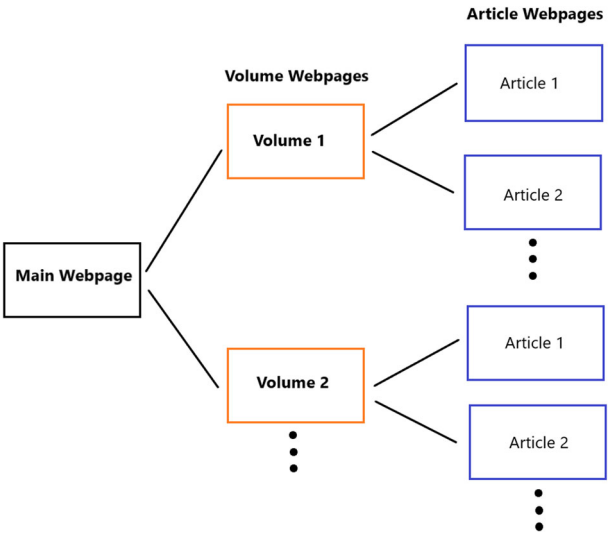


**Fig. 1** Depiction of the webpage structure for the CSDA journal for articles published since 2013

five years of data (initially) and two journals was considered to be adequate for the allocated time for this project.

For each of the journals, the web scraping process started on a main content overview webpage that includes hyperlinks to specific journal volume (and issue) 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. The COST journal uses a webpage structure where the main content webpage contains an overview of all volumes and issues published since 1999. This webpage hyperlinks to webpages that provide an overview of all published articles in an issue. These issue overview webpages further branch into individual article webpages with one webpage for each published article. The CSDA journal content overview webpage structure matches that of COST except that the main webpage for CSDA only contains volume information since 2013. CSDA stopped placing articles in issues at the end of 2012 but rather publishes twelve volumes each year since 2013 (and not twelve issues in one volume). For the years 1983 until 2012, the historic volume and issue structure can still be found on the CSDA main content overview webpage. The webpage structure for CSDA (for articles published since 2013) is shown in Fig. 1.

The first step in the web scraping process was to gather a list of hyperlinks from the main content overview webpage to access the volume webpages for the years 2015–2019 (and later for 2020). After gathering the hyperlinks from the main content overview webpage, we took each hyperlink from the first list and created a second list of hyperlinks to access the article webpages. Once we had a final list of article hyperlinks, we were able to collect the information specified above.

Each webpage is created using HTML coding. The 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 other information we intended to collect. We used regular expressions to extract only the relevant information from the article webpages. Once we were able to collect the relevant information, we compiled and stored the data into a concise table that was used for further steps of our analyses.

We also needed to gather information regarding the number of countries and the total population for 2019 and 2020. The data was not only gathered for Africa and Latin America, but 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 the IASC should move forward with the formation of an African regional section. Specifically, we compared population counts to provide insight on the activity levels we would expect to see in Africa compared to Latin America and the other continents. The Statistics Times (StatisticsTimes.com 2020) provided a table generated from the World Population Prospects 2019 (United Nations, Department of Economic and Social Affairs, Population Division 2019). The table from The Statistics Times contains data from all countries around the world, their 2019 and 2020 population, and their respective continent. For this project, we used web scraping techniques to gather this population information grouped by continent and summarized the total number of countries and population count for each individual continent. For the purpose of the group project, 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 journal web pages, we noticed that some of the country names pulled for various authors were not the official country names. Therefore, to make sure we accurately represented 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/).

### 2.3 Ranking and weighting approaches

Journal articles have different numbers of authors. Some articles are single-authored. Others have two, three, or more co-authors. Each author can list one or more affiliations. These affiliations can be in different countries and even in different geographic regions. In fact, some authors were listed with dual affiliations, e.g., in Egypt and the United States or Ecuador and Belgium. Some authors even listed three affiliations. However, less than 12% of all authors listed more than one affiliation. Depending on how authors and affiliations (and page numbers) are counted and split among multiple authors and affiliations, different author counts, page number counts, and rankings may be obtained.

Three of the five student groups initially were asked to find standardizations and visualizations of the raw and modified counts that would support the forming of a new African regional section. The two remaining student groups were asked to find evidence that would hinder the forming of a new African regional section. For the group

project, students had to web scrape the population data as part of the assignment. Students were also provided with the links from Sect. 2.3.1 via a posting in the discussion section of the online course system (Canvas), but no further instructions were provided how to translate this information from Olympic medal tables into different ranking methods for author counts, first author counts, and page counts. For this article, we translated that information as described in Sect. 2.3.2.

### 2.3.1 Meddling with Olympic medal tables

Parts of the group project were motivated by different ranking and weighting approaches, in particular for Olympic medal tables. The examples below are based on the 2016 Olympic Games that were held in Rio de Janeiro in Brazil. ESPN, an American cable sports channel, and many other outlets in the United States use a medal table ranking that is based on the total number of medals, see https://www.espn.com/olympics/summer/2016/medals. Based on this ranking, the United States "won" with 121 medals, ahead of China (70 medals) and Great Britain (67 medals). New Zealand is ranked in 14th place (18 medals).

Surprisingly, the British Broadcasting Corporation (BBC) in Great Britain used this headline on August 21, 2016: *"Rio Olympics 2016: Team GB beat China to finish second in medal table"*, see https://www.bbc.com/sport/olympics/37085511. The BBC medal table ranking is primarily based on gold medals won. Silver medals (and bronze medals) only serve as a tie-breaker in case countries won the same number of gold (and silver) medals. This is the common medal table ranking used in many European countries. The United States are still ranked 1st (with 46 gold medals), but Great Britain (27 gold medals) and China (26 gold medals) swap ranks. New Zealand now ranks 19th with four gold medals, nine silver medals, and five bronze medals (18 medals in total), ahead of Canada with four gold medals, but only three silver medals and 15 bronze medals (22 medals in total).

In the business community, it is common to adjust the medal table ranking by the Gross Domestic Product (GDP) or by the population of a country. Business Insider, an American media company, published these adjusted medal table rankings: https://www.businessinsider.com/gdp-adjusted-olympic-medal-count-2016-8 and https://www.businessinsider.com/2016-rio-olympics-medals-per-capita-2016-8. Grenada, Jamaica, and the Bahamas rank first, based on the GDP adjustment, and Grenada, Jamaica, and New Zealand rank first, based on the population adjustment. The United States, China, and Great Britain are not ranked in the top-10, based on these two adjustments. Not surprisingly, even Statistics New Zealand (Stats NZ), New Zealand's official data agency, used this ranking to showcase that New Zealand ranked third. While the original web page at http://archive.stats.govt.nz/browse_for_stats/population/estimates_and_projections/olympics-2016.aspx is no longer available, a copy of this webpage can be requested from Emailinfo@stats.govt.nz.

Further adjustments of the medal table ranking have been made by team size, see http://www.medalspercapita.com/#medals-by-team-size:2016
with Tajikistan (7 athletes), Kuwait (9 athletes), and the United States (554 athletes) ranking first. The http://www.medalspercapita.com/ web site, maintained by Craig Nevill-Manning, allows a user to try several other standardizations. One has to select

a certain Olympic, such as "2016 Rio de Janeiro," in the choice menu on the left and then try some of the listed ranking options.

The Topend Sports page at https://www.topendsports.com/events/summer/medal-tally/rankings.htm, maintained by Robert Wood, lists a few additional weighting options, in particular ways to award points to gold, silver, and bronze medals, e.g., 3–2–1 or 5–3–1, and then use the sum of points for each country to create final medal rankings. Other options that could be used for rankings could be based on the (weighted) count of the total number of medals won, i.e., a single athlete counts as one medal, a team of two players counts as two medals (e.g., in tennis doubles), and a team of 22 players (such as in soccer) counts as 22 medals. Based on this ranking, a country would have to win in team sports to be highly ranked overall according to this adjustment. Even others suggested to award points for places 1 through 4 or even for places 1 through 10 and not only for places 1 through 3 and then do an extended ranking based on these points.

Sommers (1996, 1997), Morton (2002), Moosa and Smith (2004), and Wang et al. (2011) are among those who provided examples on how to look at Olympic medal table rankings from various statistical perspectives.

### 2.3.2 Meddling with author rankings

A natural first attempt that closely followed the population-based adjustment for the Olympic medal table rankings was done to adjust author and page counts with respect to the population in Latin America and Africa, but that did not yield any meaningful results as shown in Sect. 3.

Next, several weighting methods were attempted to see whether there would be a difference in the ranking order of Africa and Latin America in regards to author count and page count. Just considering the impact of the first authors resembles Olympic medal table rankings that are primarily based on Gold medals. For a journal article, the assumption could be made that first authors often lead the research presented in an article and, therefore, first authors deserve most (or even all) the weight. The reason for attempting these different weighting methods was to determine whether we were counting and portraying author and affiliation counts in a meaningful way or whether a different weighting method might change the ranking order. The description of each weighting method is given below. Some motivating examples are shown in Table 1.

(A) Count all authors and affiliations with weight 1 (i.e., an author with three affiliations gets counted three times). The sum of this count is equal to the total number of affiliations for each article.
(B) Count all authors and affiliations with weight $1/m$ for $m$ affiliations overall. The sum of this count is equal to one for each article.
(C) Count all authors and affiliations with weight $1/n$ for $n$ co-authors (i.e., an article with three co-authors gets counted $1/3$ times for each co-author (for each affiliation)). The sum of this count tends to be greater than one for many articles, except for articles where each author has only one affiliation.
(D) Count all authors and affiliations with weight $1/n$ for $n$ co-authors and weight $1/m_i$ for $m_i$ affiliations of author $i$ (i.e., an article with three co-authors gets

counted $1/3$ times for each co-author—this is then further adjusted by the number of affiliations (i.e., an author with three affiliations gets counted $1/3 \cdot 1/3 = 1/9$ for each affiliation)). The sum of this count is equal to one for each article.

(E) Count first authors and affiliations with weight 1. The sum of this count is equal to the number of affiliations of the first author for each article.

(F) Count first authors and affiliations with weight $1/m_i$ for $m_i$ affiliations of the first author. The sum of this count is equal to one for each article.

Similar methods as above were used to count the number of pages for authors in Latin America and Africa. For those page weights, the page numbers for each article were multiplied by the previously obtained author weights. The raw number of pages per article was obtained by taking the end page number of an article and subtracting the start page number of an article, then adding one.

For CSDA articles in 2020 that do not list page numbers, but only article numbers, the page count for each article was obtained by downloading each article onto a local device. We then used the *pdftools* R package to calculate the page length of each article. We took the results and added it to our existing data. Weighting methods (A) and (E) were used in Medri et al. (2021). In Table 1, a sample of three articles from the COST journal were selected to display how the weights were calculated according to all six different weighting methods.

### 2.4 Data visualization

To compare the African and Latin American activity levels, we narrowed our focus to the factors from Table 1, aggregated separately for the years 2015–2020. In addition, we compare Africa and Latin America to other geographic regions to show counts in context. For others, we focused on Africa and Latin America as our two primary geographic regions of interest. Our visualizations primarily make use of bar charts and line charts. In addition, we also compared the 2019 and 2020 population among all the continents to gain an understanding of how Africa compares to Latin America, Asia, Europe, and Oceania for further insight about the demographics of this geographic region.

## 3 Results

Our analyses in this section focus on the information gathered from the web scraping process for the years 2015–2020. We gathered information from 464 articles from the COST journal and 1069 articles from the CSDA journal. Our web scraping procedure effectively accessed 99.87% of the total 1533 articles available. Information for the two articles (0.13%) that were not accessed automatically via web scraping (because of a system error) was manually entered at the end of the web scraping process. While we gathered data on all geographic regions, we focused our efforts on comparing the African and Latin American regions. Table 2 shows a summary of those web scraping results. In this table, 'Articles' include all types of articles such as original papers, short notes, and editorials—with the exception of erratums. 'Authors' include

**Table 1** Author counts and page counts for six different weighting methods (A) through (F)

| Author name | Country | All authors | | | | First authors | | Pages all | | | | Pages first | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | (A) | (B) | (C) | (D) | (E) | (F) | (A) | (B) | (C) | (D) | (E) | (F) |
| Pereira and de Andrade (2015): 2 authors, 2 affiliations, 20 pages | | | | | | | | | | | | | |
| André G. C. Pereira | Brazil | 1 | 1/2 | 1/2 | 1/2 · 1 | 1 | 1 | 20 · 1 | 20 · 1/2 | 20 · 1/2 | 20 · 1/2 · 1 | 20 · 1 | 20 · 1 |
| Bernardo B. de Andrade | Brazil | 1 | 1/2 | 1/2 | 1/2 · 1 | 0 | 0 | 20 · 1 | 20 · 1/2 | 20 · 1/2 | 20 · 1/2 · 1 | 20 · 0 | 20 · 0 |
| | Sum: Brazil | 2 | 1 | 1 | 1 | 1 | 1 | 40 | 20 | 20 | 20 | 20 | 20 |
| | Grand Sum | 2 | 1 | 1 | 1 | 1 | 1 | 40 | 20 | 20 | 20 | 20 | 20 |
| Balakrishnan and Pal (2015): 2 authors, 3 affiliations, 39 pages | | | | | | | | | | | | | |
| N. Balakrishnan | Canada | 1 | 1/3 | 1/2 | 1/2 · 1 | 1 | 1 | 39 · 1 | 39 · 1/3 | 39 · 1/2 | 39 · 1/2 · 1 | 39 · 1 | 39 · 1 |
| Suvra Pal | Canada | 1 | 1/3 | 1/2 | 1/2 · 1/2 | 0 | 0 | 39 · 1 | 39 · 1/3 | 39 · 1/2 | 39 · 1/2 · 1/2 | 39 · 0 | 39 · 0 |
| Suvra Pal | South Africa | 1 | 1/3 | 1/2 | 1/2 · 1/2 | 0 | 0 | 39 · 1 | 39 · 1/3 | 39 · 1/2 | 39 · 1/2 · 1/2 | 39 · 0 | 39 · 0 |
| | Sum: Canada | 2 | 2/3 | 2/2 | 3/4 | 1 | 1 | 78 | 26 | 39 | 29.25 | 39 | 39 |
| | Sum: South Africa | 1 | 1/3 | 1/2 | 1/4 | 0 | 0 | 39 | 13 | 19.5 | 9.75 | 0 | 0 |
| | Grand Sum | 3 | 1 | 3/2 | 1 | 1 | 1 | 117 | 39 | 58.5 | 39 | 39 | 39 |
| Kolbe et al. (2015): 4 authors, 6 affiliations, 24 pages | | | | | | | | | | | | | |
| Jens Kolbe | Germany | 1 | 1/6 | 1/4 | 1/4 · 1/2 | 1 | 1/2 | 24 · 1 | 24 · 1/6 | 24 · 1/4 | 24 · 1/4 · 1/2 | 24 · 1 | 24 · 1/2 |
| Jens Kolbe | Germany | 1 | 1/6 | 1/4 | 1/4 · 1/2 | 1 | 1/2 | 24 · 1 | 24 · 1/6 | 24 · 1/4 | 24 · 1/4 · 1/2 | 24 · 1 | 24 · 1/2 |
| Rainer Schulz | United Kingdom | 1 | 1/6 | 1/4 | 1/4 · 1 | 0 | 0 | 24 · 1 | 24 · 1/6 | 24 · 1/4 | 24 · 1/4 · 1 | 24 · 0 | 24 · 0 |
| Martin Wersing | United Kingdom | 1 | 1/6 | 1/4 | 1/4 · 1 | 0 | 0 | 24 · 1 | 24 · 1/6 | 24 · 1/4 | 24 · 1/4 · 1 | 24 · 0 | 24 · 0 |
| Axel Werwatz | Germany | 1 | 1/6 | 1/4 | 1/4 · 1/2 | 0 | 0 | 24 · 1 | 24 · 1/6 | 24 · 1/4 | 24 · 1/4 · 1/2 | 24 · 0 | 24 · 0 |
| Axel Werwatz | Germany | 1 | 1/6 | 1/4 | 1/4 · 1/2 | 0 | 0 | 24 · 1 | 24 · 1/6 | 24 · 1/4 | 24 · 1/4 · 1/2 | 24 · 0 | 24 · 0 |
| | Sum: Germany | 4 | 4/6 | 4/4 | 4/8 | 2 | 1 | 96 | 16 | 24 | 12 | 48 | 24 |
| | Sum: United Kingdom | 2 | 2/6 | 2/4 | 4/8 | 0 | 0 | 48 | 8 | 12 | 12 | 0 | 0 |
| | Grand Sum | 6 | 1 | 3/2 | 1 | 2 | 1 | 144 | 24 | 36 | 24 | 48 | 24 |

Columns (A) through (D) relate to all authors and columns (E) and (F) relate to first authors only, giving all additional authors a weight of zero

**Table 2** Number of articles, number of authors, and number of pages published in COST and CSDA, based on our web scraping results

| Year | COST | | | CSDA | | |
|---|---|---|---|---|---|---|
| | Articles | Authors | Pages | Articles | Authors | Pages |
| 2015 | 60 | 138 | 1248 | 144 | 381 | 1886 |
| 2016 | 75 | 192 | 1619 | 255 | 637 | 3545 |
| 2017 | 75 | 184 | 1732 | 188 | 512 | 2593 |
| 2018 | 82 | 214 | 1919 | 160 | 446 | 2352 |
| 2019 | 82 | 231 | 1867 | 154 | 409 | 2260 |
| 2020 | 90 | 231 | 2047 | 168 | 478 | 2800 |
| Total | 464 | 1190 | 10,432 | 1069 | 2863 | 15,436 |

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' represent 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.

Using the information from Table 2, we created several visualizations to help us compare the African geographic region with the Latin American geographic region. When comparing these two regions, we made some general observations as to whether or not Africa is ready for its own regional IASC section from a publication perspective.

Figure 2 shows that the population in Africa in 2020 ($\approx$ 1.37 billion) was about twice as large as the population in Latin America ($\approx$ 0.66 billion). The same pattern was seen when evaluating population counts for 2019. As seen, Africa is the second largest continent following Asia regarding population size. When comparing the African population with population counts in Asia, Europe, and Latin America which all contain their own regional IASC section, it appears that the population in Africa is large enough to produce a sufficiently large output of publications in the field of computational statistics (as published in COST and CSDA).

While we would expect to see proportions of authors from Africa and Latin America that are comparable to the general population proportions shown in Fig. 2, Figure 3 shows that the author to population ratio is about five times greater in Latin America than in Africa in 2019. Looking at the author to population ratio for 2020, we noticed a slight decrease in both Africa and Latin America, with the ratio for Latin America roughly being four times greater than for Africa. The author counts are based on weighting method (A) described in Sect. 2.3.2.

Figure 4 depicts that the proportion of pages published in Latin America is also much greater than the proportion of pages published in Africa when compared to the total population in these two geographic regions in 2019 and 2020. It is important to note that even though Africa has a larger population count than Latin America, when measuring against the relevant ratio of authors and pages, Africa has a much smaller ratio. For this reason, we focused on pure author and page counts for all

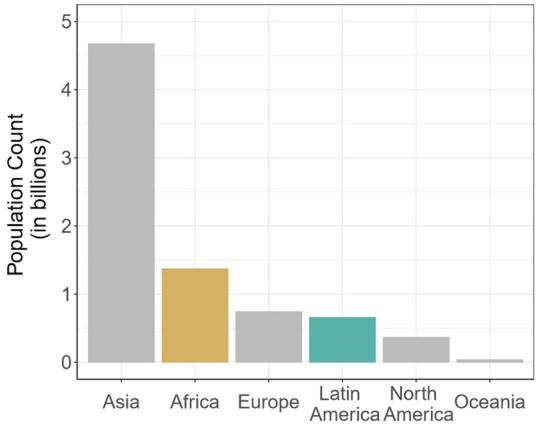


**Fig. 2** 2020 world population count (in billions) by geographic region. As previously defined, North America is defined as the United States and Canada only while all other countries on the American continent belong to Latin America for the purpose of this article

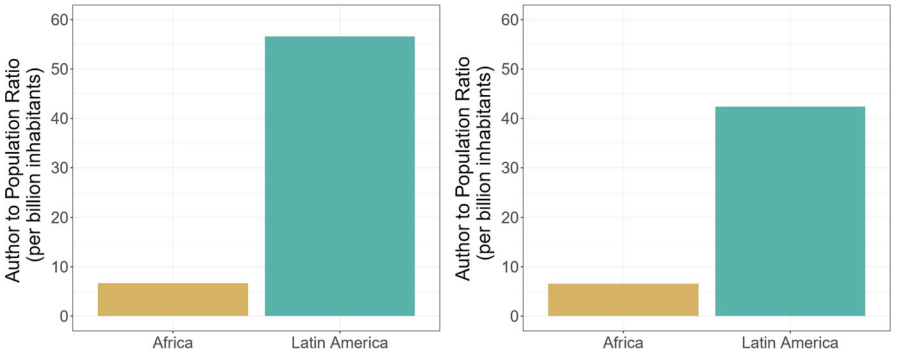


**Fig. 3** Number of authors to population ratio in Africa and Latin America in 2019 (left) and 2020 (right)

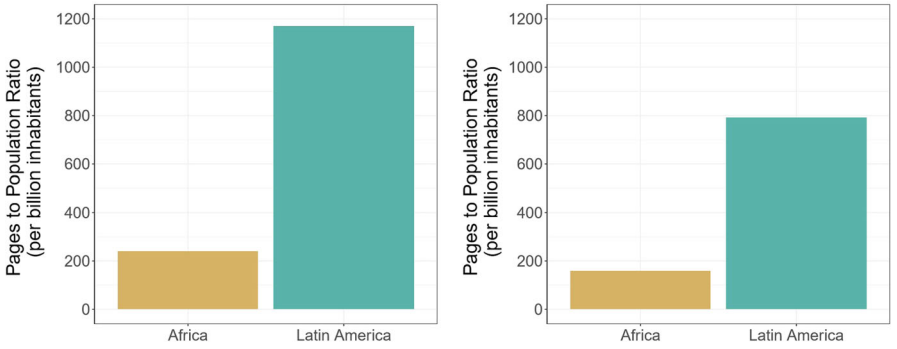


**Fig. 4** Number of pages to population ratio in Africa and Latin America in 2019 (left) and 2020 (right)

remaining results without any further adjustment with respect to the population in a given geographic region.

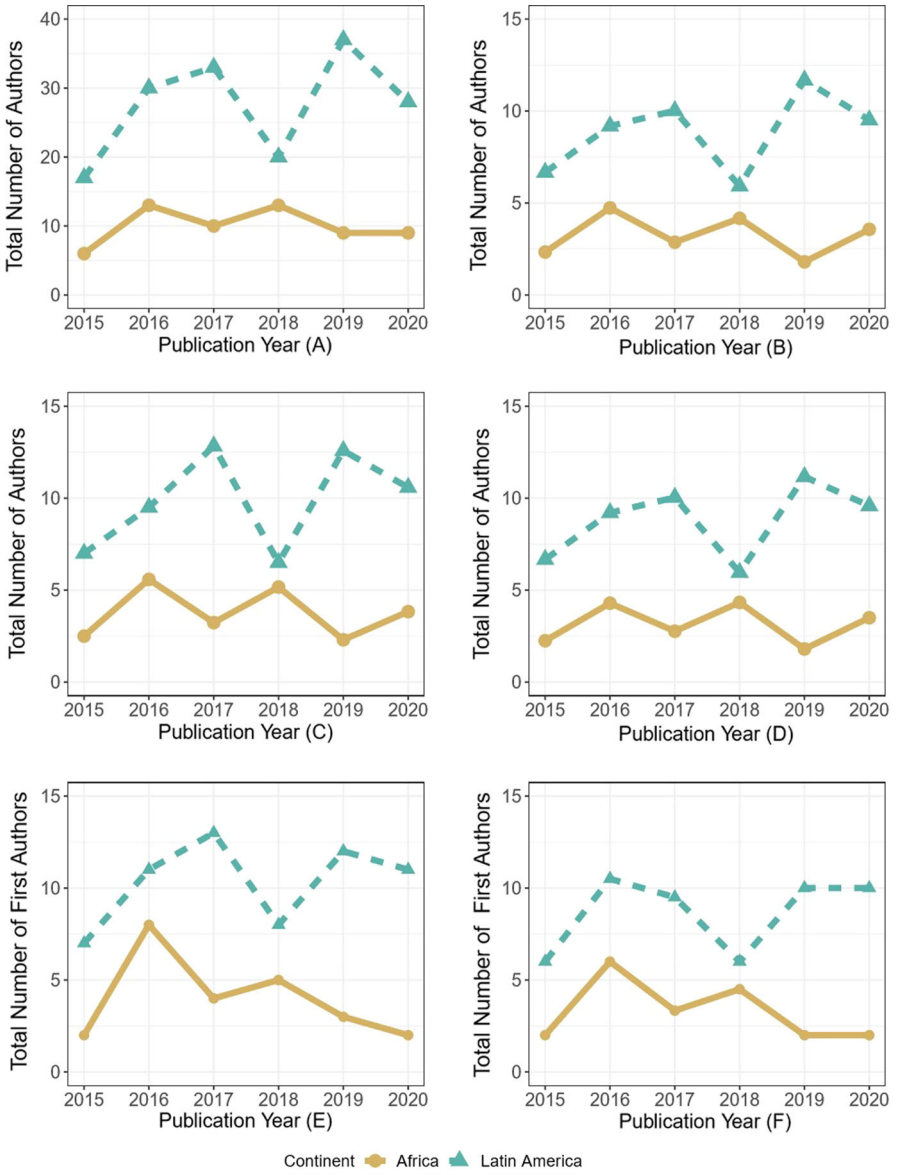


**Fig. 5** Results of several weighting methods on the number of author counts (A–D) and first author counts (E–F). See text for details

The results of the six weighting methods discussed in Sect. 2.3.2 are shown in Fig. 5 for author counts and Fig. 6 for page counts. Fig. 7 shows the results of weighting methods (E) and (F) for the number of first author counts for all geographic regions.

From what we can see in Fig. 5, regardless of the weighting method used, the results for all six weighting methods show similar trends. Latin America consistently shows

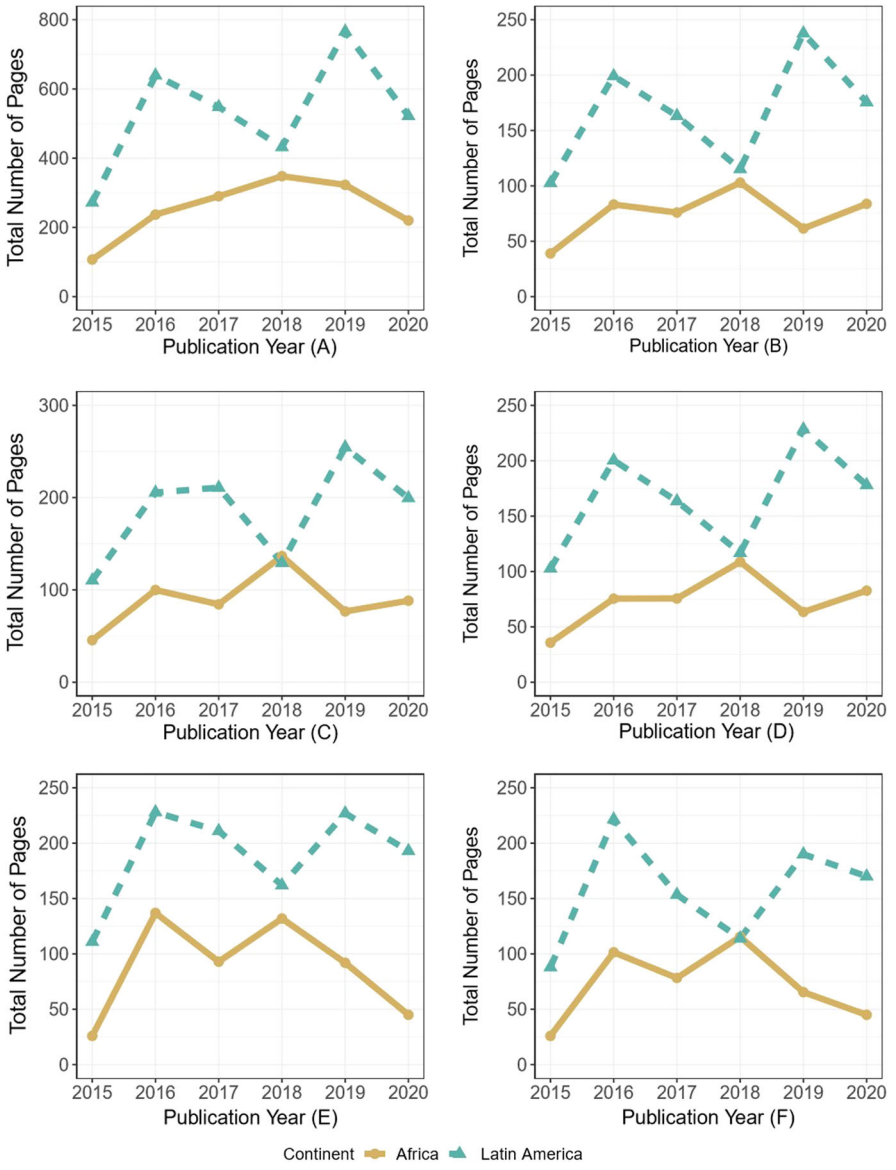


**Fig. 6** Results of several weighting methods on the number of page counts for all authors (A–D) and page counts for first authors (E–F). See text for details

higher author counts for all 6 years for all six weighting methods. Overall, there is a slight positive trend for Latin America for all six author counts. For Africa, there is at best a minimal positive trend for author counts (A) through (D) and possibly even a negative trend for first author counts (E) and (F). There is a considerable annual variation for each geographic region. For Latin America, the year 2018 looks rather

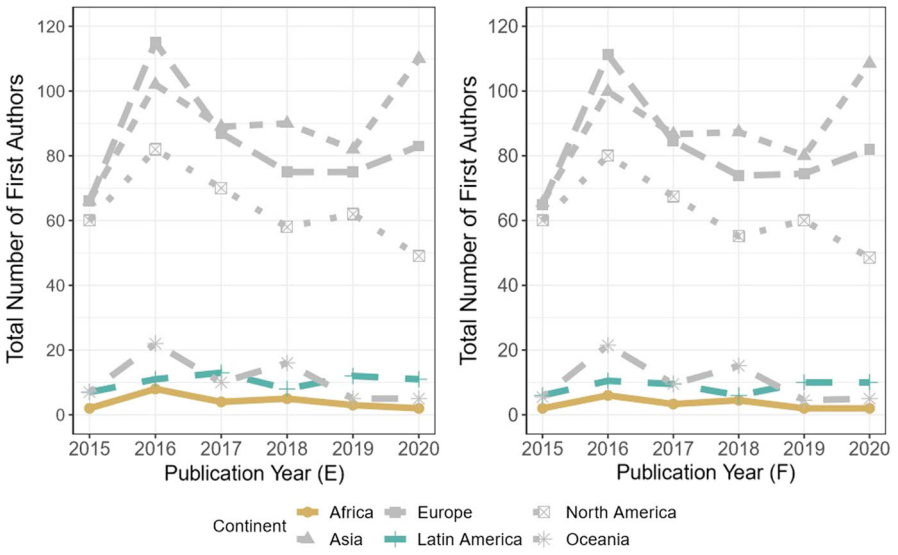


**Fig. 7** Results of weighting methods (E) and (F) on the number of first author counts for all geographic regions. See text for details

unusual with considerably less author counts for all six weighting methods, compared to 2017 and 2019. In fact, Latin America and Africa have almost identical author counts for weighting methods (B), (C), (D), and (F) in 2018.

In Fig. 6, the results for the page counts for all six weighting methods again are all very similar and show similar trends. Much like what was observed with the author counts (in Fig. 5), Latin America consistently shows higher page counts for all six years for all six weighting methods. Overall, there is a slight positive trend for Latin America and also for Africa for all six page counts. The slope for Latin America seems to be steeper than for Africa, but no formal tests have been performed as this was not part of the original assigned project. In addition, there is considerable annual variation for each geographic region. For Latin America, the year 2018 again looks rather unusual with a lower number of page counts for all six weighting methods, compared to 2017 and 2019. Interestingly, Africa had page counts in 2018 for all six weighting methods that roughly equaled the page counts for Latin America in 2015, i.e., before the Latin American regional IASC section was founded. However, this was a local maximum as all page counts for Africa decreased for 2019 and 2020.

Finally, Fig. 7 is an extension of Fig. 5 for methods (E) and (F) that shows the first author counts in Africa and Latin America in the global context of all geographic regions. While Figs. 5 and 6 suggest that Latin America is ahead of Africa with respect to most of the weighting methods in most years, Fig. 7 suggests that both Latin America and Africa considerably trail Asia, Europe, and North America. Latin America and Africa have similar first author counts as Oceania that has a much smaller population as shown in Fig. 2. Similar patterns (not shown here) can be seen for weighting methods (A–D) and for page counts.

## 4 Students Observations and Feedback

In addition to the quantitative results based on the data obtained from web scraping, we were interested in gathering student perspectives of the "Data Technologies" group project. For more details on this course and the complete description of this group project, visit https://math.usu.edu/~symanzik/teaching/2019_stat5080/stat5080.html. Additionally, the description of the group project can be found in Appendix B We identified questions that would give relevant and honest answers about the assignment that would benefit any instructor interested in assigning similarly complex assignments in their courses. All 24 students from the course were invited to provide their feedback via e-mail to the representative from their group who participated in the ongoing work after the end of the course. Typically, two or three students from each group (out of four or five group members) provided some feedback. The five representatives from the groups aggregated the individual responses from their group members. The summary below is based on these group responses and not on the individual student responses.

- 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.?

  Most groups only met with all members two or three times. However, these groups either formed subgroups or assigned the tasks to individual group members. The group leader for most groups combined the individual or subgroup results and R code. One group met as a whole two or three times each week and worked collectively on all tasks.
- How did you arrange the writeup for this assignment?
  In most groups, one student took the lead. Other students contributed shorter text parts, graphs, tables, etc., based on the work they did. They also proofread and edited the writeup. One group prepared the final writeup in the last collective group meeting.
- Which tools did you use, e.g., R, RStudio, box, dropbox, github, Overleaf (an online collaborative LaTeX editor), Google Docs, etc.?
  All groups used R and RStudio and communicated via regular e-mail. Three of the five groups also used GoogleDocs. In addition, Box, Dropbox, and Excel were used by other groups.
- 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?
  Some groups used an alternative website for the country population. Other groups changed the order in which some of the tasks were completed. Other than these deviations, the groups followed the instructions (and order) relatively closely.
- How relevant was this assignment for your overall "Data Technologies" learning experience?

  Most groups described this experience as "extremely relevant" and "very helpful," in particular to go through all necessary steps from data collection, data cleaning, data analysis, and visualization — something similar to what could be found in a future workplace. However, one student had a somewhat negative impression

describing the assignment as a "grab-bag of all kinds of things" even though this is a good description of data technologies and data science.

- What was particularly useful/interesting in this assignment?
  Most groups agreed that web scraping was particularly interesting. Other students liked the "open-endedness" and that there were no pre-determined "correct" answers. Another common response was how students appreciated learning how to overcome unexpected problems.
- What was not useful/interesting at all in this assignment?
  It was a consensus that extra time would have helped. As a course with undergraduate and graduate students, some of the undergraduate students felt that the project was "a little bit of a stretch".
- For those of you who have started a job since the end of the "Data Technologies" course, does any of the material/knowledge from this assignment help in your current job? Which material?
  Out of four students to whom this question applied, two mentioned that knowledge about regular expressions has been helpful in their current work. One of these students added that the overall project experience has been beneficial. Two students accepted jobs that did not relate to this material. The remaining students from the course who answered the survey had not started a job at that time or did not answer this question.
- What, if any, challenges did you face working with a group on an assignment of this magnitude?
  Dealing with time constraints (e.g., when to meet as a group) was a common challenge. Some groups mentioned that not all group members were equally interested and engaged in the project. In one group, the graduate student "leader" was unable to fill this leadership role. In another group, one student dropped the course requiring that group to reassign the tasks to the four remaining group members. Finally, some groups experienced some challenges when combining individual R code segments.
- 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)?

  The median estimate was about 70 hours (roughly 14 hours per student per group). The minimum estimate was 30 to 50 hours and the maximum estimate was 90 hours. Most groups acknowledged that even the approximate amount of time spent would be difficult to estimate accurately.
- Any additional comments?
  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." One student indicated: "The assignment was definitely a lot of fun and I would recommend assignments like these for future courses. It taught me a lot about using relevant information to draw conclusions and make statements."

## 5 Conclusions and outlook

As stated in the introduction, this article has a dual purpose: In the first part, we try to answer the question whether Africa is ready for a new regional IASC section, based on publication records in two major statistical computing journals. In the second part, we look at student perspectives and student feedback for the underlying group project.

A conclusion for part one is that, based on Figs. 5 and 6, Africa's activity levels with respect to author counts and page counts for articles published in COST and CSDA have somewhat increased from 2015 to 2020 for some of the six weighting methods that were investigated. This holds more for the number of page counts than for the number of author counts. Independently from the weighting method, Africa almost consistently trails Latin America. However, to address whether Africa is ready for a new regional IASC section, one should not compare current numbers for Africa and Latin America, but rather compare current numbers for Africa with numbers from Latin America, in particular in 2015, i.e., the year before IASC-LARS was founded. Based on such a comparison, Africa looked favorably, in particular in 2018, for most of the weighting methods for the number of page counts. For some of the weighting methods for the number of page counts, Africa also looked somewhat favorable for isolated years when compared with Latin America's counts for 2015, e.g., for the total number of pages for first authors according to method (E) in 2016. Both regions experienced considerable decreases in 2020. It is easy to speculate that this is due to the Coronavirus (COVID-19) pandemic.

When we compare the African region with the Latin American region, we find that the rates of increase from 2015 to 2020 seem to be greater in Latin America than in Africa (although no formal statistical tests were conducted) for most of the weighting methods. While it is possible that the creation of the new regional IASC section created in Latin America in late 2016 contributed to the rate of increase, our analysis is exploratory and doesn't examine causal relationships. 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, i.e., before the foundation of IASC-LARS. Moreover, when comparing Africa and Latin America globally with Asia and Europe (as shown in Fig. 7), both of which have regional IASC sections, it is hard to justify why Latin America should have a regional IASC section and Africa should not have such a section as the differences between both regions are minimal in the global context.

As a conclusion for part two, we observe that while such a group project is unconventional, students appeared to embrace this type of challenge and recognize the applicability and relevance with a generally positive outlook. If managed well, instructors who teach data technology methods and similar courses may provide a more meaningful learning experience to their students if they assign similar projects. However, it is essential to carefully monitor the timing of the project and possibly shorten the number of sub-tasks or extend some deadlines if necessary, e.g., for final presentations and the submission of the project. Also, it is necessary to ensure that weaker students can still contribute to the overall project. This was the reason why this project

was designed as a group assignment where each group member can take on tasks that match their skills.

As an epilogue, it should be noted that there are new favorable developments in Africa since the conclusion of the project and data gathering. Based on an internal IASC membership report, Africa had 33 IASC members in August 2021 (almost twice as many as in August 2020) which by far bypassed the minimum number of 20 members to form a regional IASC section. Moreover, several IASC members in Nigeria formed a regional *IASC African Members Group* towards the end of 2020 (see https://iasc-isi.org/2021/02/17/existence-of-international-association-of-statistical-computing-iasc-african-members-group/) that can be seen as a predecessor of a regional section. It is not difficult to imagine that there will be a regional African IASC section (IASC-AFRS) in the next few years.

**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 Sect. 4.

**Funding** No funding was received to assist with the preparation of this manuscript.

## Appendices

### A Appendix: R Tools

To complete the data manipulations and visualizations, we made use of the following `R` packages: "boxr"(Rocks et al. 2019), "cowplot"(Wilke 2019), "data.table"(Dowle and Srinivasan 2019), "dplyr"(Wickham et al. 2020), "ggplot2"(Wickham 2016), "gridExtra" (Auguie 2017), "httr"(Wickham 2019), "janitor"(Firke 2020), "kableExtra"(Zhu 2019), "pdftools"(Ooms 2021), "sqldf"(Grothendieck 2017), "tibble"(Müller and Wickham 2020), "XML" (Temple Lang 2020), and "xtable"(Dahl et al. 2019).

### B Appendix: STAT 5080/6080 Data Technologies Homework Assignment Instructions

1 (84 Points) You are asked to help me in my current role as the President of the *International Association for Statistical Computing* (IASC) with a simple question: How much activity is there in Africa with respect to computational statistics?
**Background**: The 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 http://iasc-isi.org/ for more details about the IASC—new students members are always welcome!). The IASC currently has three regional sections in Europe (IASC-ERS), in Asia (IASC-ARS), and in Latin America (IASC-LARS), that 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. To establish a new regional section, there must be a minimum number of IASC members in that geographic

region. Moreover, the IASC General Assembly (GA) must approve a new regional section. That approval likely depends on the question whether the new 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. This leads to the question whether there is currently enough high-level activity in Africa with respect to computational statistics.

**Approach:** 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 web pages from university and research institutes. In this HW, we will use author information from two leading journals in the field of computational statistics to answer this question. Specifically, we will extract author information from authors located in Africa in the past few years and see how this relates to similar author information from authors located in Latin America from a few years ago.

You will be asked to create many tables for this HW. See how to use the *kable* function from the *knitr & kableExtra* R packages (easier, but less powerful) or the *xtable* function from the *xtable* R package (harder, but more powerful) to export your tabular data and data frames from R into a nice printable table in your LaTeX document.

This homework is a group homework with groups of 4 or 5 members. Groups were assigned in class on Tuesday 11/5/2019. The group head has to submit the answers for the following questions on behalf of the entire group. The other group members do not have to submit answers separately. Each group member will obtain the same score.

As always, make sure to include your R part and a resulting graph if the question asks for a graph. Use *tidyverse* functionality whenever possible.

a (1 Point) Load all required R packages to answer this question. Show your R code.

b (15 Points) Download a list of African countries from http://statisticstimes.com/demographics/african-countries-by-population.php. This contains a nice table. Extract the country names and the most recent population count (from 2018) from this table and transform these numbers into a numeric. Write this information into an external csv file with just these two columns.
Repeat these steps for the five other listed continents at the bottom of the Africa page, i.e., for Asia, Europe, North America, South America, and Oceania. Further adjust the countries for North and South America: Only the United States and Canada belong to North America for the purpose of this HW. We consider Mexico and everything further to the south as South (Latin) America. You should manually check your results. Do you read in all countries correctly, do you extract the correct population counts and are those numeric, etc.?
Produce a final summary table for this question part that contains six rows (one for each continent) and three columns: The name of the continent, the number of countries (only two for North America) in that continent, and the total 2018 population for this continent. Include your R code and this final table in your HW.

c (24 Points) Extract author information from the *Springer* journal *Computational Statistics* (COST) for the past 5 years: Start at https://link.springer.com/journal/volumesAndIssues/180 and extract information, beginning with Volume 30, Issue 1, March 2015, and ending with Volume 34, Issue 4, December 2019. Do not continue for 2020, even if the next issue(s) get posted while you work on this HW.
For each published article in each issue, obtain the following information and store in a data frame:
- Journal (COST, CSDA).
- Year.
- Volume.
- Issue.
- Title of the article.
- Number of authors for the article.
- Author name.
- Author affiliation.
- Author country.
- Author order (1, . . ., number of authors for the article).
- Start page of article.
- End page of article.

Note that most articles have more than just one author. As this is basically a long format, we repeat similar information for each of the authors of an article. Manually check that your results are meaningful, that you capture information for all authors, etc. Be careful with special issues, editorials, and invited/discussion papers. The list of articles in an issue may extend beyond one web page. We exclude the *Taylor & Francis* journal *Journal of Computational and Graphical Statistics* (JCGS) from this HW as they use at least three different ways to store author affiliations. Based on what I have sampled, COST seems to be consistent, but no guarantees. So, check that you do not just get NAs for some issues.
Overall, you should obtain author information for five years. Once more, check your result by sampling a few articles and checking that the information from those articles has been correctly stored in your data frame.
Repeat, now for the *Elsevier* journal *Computational Statistics & Data Analysis* (CSDA) for the past 5 years: Start at https://www.sciencedirect.com/journal/computational-statistics-and-data-analysis/issues, beginning with Volume 81, January 2015, and ending with Volume 140, December 2019.
Create a similar data frame as for COST. **Do not** include any of the forthcoming issues for 2020.
Write your two data frames into two external csv files, one for each journal. Produce a final summary table for this question part that contains five rows (one for each year) and five columns: year, total number of articles in COST in the year, total number of authors in COST in the year (some authors may appear in more than one article; if so, count them multiple times), total number of articles in CSDA in the year, and total number of authors in CSDA in the

year (some authors may appear in more than one article; if so, count them multiple times). Include your R code and this final table in your HW.

d (8 Points) Likely, the country names from part (b) will not entirely match the author countries from part (c). Identify non-matching author countries from part (c) and write R code that adjusts these author countries to the format used in part (b). Update the names in your data frames and write to two new external csv files. Do not overwrite your previously created external files.
Likely problems with country names could be different versions for the same country, e.g., United States, USA, U.S.A., etc., instead of United States of America, different spellings (e.g., Viet Nam or Vietnam), use of special characters (e.g., Côte d'Ivoire), prefixes and postfixes to a country name (e.g., Republic of Moldova or just Moldova; Russian Federation or just Russia), name changes (e.g., TFYR Macedonia, Macedonia, or North Macedonia), and more.
Be careful when you work with regular expressions and what you match: There exist Niger and Nigeria (two different countries), Congo and Democratic Republic of the Congo (two different countries), Sudan and South Sudan (two different countries), Dem. People's Republic of Korea and Republic of Korea (two different countries; often referred to as North Korea and South Korea for simplicity), and many more similar examples.
Create a summary table that lists the different versions found in part (c) and to which country name from part (b) they get translated. There may be multiple versions from part (c) that all get translated to the same version from part (b). Arrange this table in alphabetical order, according to the versions from part (b). Include your R code and this table in your HW.

e (16 Points) Produce the following tabular summaries. Create meaningful headings in your tables and meaningful table captions.
   i Top-20 countries, based on all authors for COST for all years combined.
   ii Top-20 countries, based on all authors for CSDA for all years combined.
   iii Top-20 countries, based on all authors for both journals for all years combined.
   iv Top-20 countries, based on first author only for COST for all years combined.
   v Top-20 countries, based on first author only for CSDA for all years combined.
   vi Top-20 countries, based on first author only for both journals for all years combined.
   vii For each of the five years, count of number of first authors and number of pages for first author for Africa and Latin America for both journals combined (two columns for each continent); plus a final row that lists the totals for each column.
   viii For each of the five years, count of number of all authors and number of pages for all authors for Africa and Latin America for both journals combined (two columns for each continent); plus a final row that lists the totals for each column. If an article has two authors, each author is counted twice and the number of pages is counted twice. If there are five

authors, the author count will be five and the number of pages is counted five times.

Include your R code and these eight tables in your HW.

f (12 Points) You have to be a devil's advocate: Groups 1, 3, and 5 should **support** the forming of a new African regional section. Groups 2 and 4 should try to **hinder** the forming of a new African regional section.

To do this from your perspective, create similar tables as in part (e), but use a different standardization, e.g., using total population in a continent, using different weights for the number of authors and pages (e.g., for a 20-page article with 2 authors, each author only counts 1/2 or only gets 10 pages; for a 20-page article with 5 authors, each author only counts 1/5 or only gets 4 pages), etc. You want to compare Africa in 2018 and 2019 with Latin America in 2015 and 2016 (before the IASC-LARS regional section was formed). Also consider time series that show an increasing trend over the five-year period to show your point of view, e.g., that Africa is "there" or "almost there" even if it has not exactly reached the numbers for Latin America in 2015 and 2016.

Create at least three meaningful graphical summaries of three different tables that support your point of view. These might be bar charts, spine plots (these might be very useful), line charts / time series, or others. If you are familiar with choropleth maps, consider creating side-by-side choropleth maps for Africa and Latin America where one shows high counts or standardized values for most of the countries and the other shows low counts or standardized values for most of the countries.

You are not allowed to apply any of the rules for bad graphs from Stat Viz I, i.e., you cannot wiggle the baseline in stacked bar charts, you cannot graph data out of context by just comparing the top country in Africa with the top country in Latin America (and omit all other countries), etc. Also, when you create related plots, they typically have to follow the small multiple principle, in particular use identical scales and colors / intervals in maps.

Include your R code and these three (or more) graphs in your HW.

g (8 Points) Provide a final summary and discussion of your results from your perspective from part (f) and argue why a regional section in Africa should be formed / should not be formed. Be formal and refer to specific tables and figures from your previous parts in this summary.

If you could not create three different graphs that support your point of view, then concede! Support the forming of a regional section in Africa even if your original task was to hinder it. Or, admit that there is nothing in the data that supports the forming of a regional section in Africa even if your original task was to support it.

This summary should be between 1/2 and 1 page in length.

## References

Auguie B (2017) gridExtra: miscellaneous functions for "Grid" graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Balakrishnan N, Pal S (2015) An EM algorithm for the estimation of parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood- and information-based methods. Comput Stat 30(1):151–189. https://doi.org/10.1007/s00180-014-0527-9

Dahl DB, Scott D, Roosen C, Magnusson A, Swinton J (2019) xtable: Export tables to LaTeX or HTML. R package version 1.8-4. https://CRAN.R-project.org/package=xtable

Dowle M, Srinivasan A (2019) data.table: Extension of 'data.frame'. R package version 1.12.8. https://CRAN.R-project.org/package=data.table

Firke S (2020) Janitor: simple tools for examining and cleaning dirty data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

Grothendieck G (2017) sqldf: manipulate R data frames using SQL. R package version 0.4-11. https://CRAN.R-project.org/package=sqldf

Hardin J, Hoerl R, Horton NJ, Nolan D, Baumer B, Hall-Holt O, Murrell P, Peng R, Roback P, Temple Lang D, Ward MD (2015) Data science in statistics curricula: preparing students to "Think with Data". Am Stati 69(4):343–353. https://doi.org/10.1080/00031305.2015.1077729

Kolbe J, Schulz R, Wersing M, Werwatz A (2015) Identifying Berlin's land value map using adaptive weights smoothing. Comput Stat 30(3):767–790. https://doi.org/10.1007/s00180-015-0559-9

Medri J, Coltrin J, Fleming A, Hilyard C, Tellez R, Symanzik J (2021) Is Africa ready for a new regional IASC section? Results and student experiences from a web-scraping assignment. In: Moncayo-Martinez LA, Munoz DF (eds) LACSC2021 Proceedings, V Latin American conference on statistical computing April 19th-21st, 2021. Sello Editorial ITAM, Mexico City, pp 81–89

Moosa IA, Smith L (2004) Economic development indicators as determinants of medal winning at the Sydney Olympics: an extreme bounds analysis. Aust Econ Pap 43(3):288–301. https://doi.org/10.1111/j.1467-8454.2004.00231.x

Morton RH (2002) Who won the Sydney 2000 Olympics? An allometric approach. J R Stat Soc Ser D 51(2):147–155. https://doi.org/10.1111/1467-9884.00307

Müller K, Wickham H (2020) tibble: simple data frames. R package version 3.0.1. https://CRAN.R-project.org/package=tibble

Munzert S, Rubba C, Meißner P, Nyhuis D (2014) Automated data collection with R: a practical guide to web scraping and text mining. Wiley, Chichester, UK

Murrell P (2009) Introduction to data technologies. Chapman and Hall, Boca Raton, FL

Ooms J (2021) pdftools: text extraction, rendering and converting of pdf documents. R package version 3.0.1. https://CRAN.R-project.org/package=pdftools

Pereira AG, de Andrade BB (2015) On the genetic algorithm with adaptive mutation rate and selected statistical applications. Comput Stat 30(1):131–150. https://doi.org/10.1007/s00180-014-0526-x

R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rocks B, Lyttle I, Day N (2019) boxr: interface for the 'Box.com API'. R package version 0.3.5. https://CRAN.R-project.org/package=boxr

Rundel C, Çetinkaya-Rundel M (2016) Taking a chance in the classroom: La Quinta is Spanish for "Next to Denny's". Chance 29(2):53–57. https://doi.org/10.1080/09332480.2016.1181966

Sommers PM (1996) Meddling with 1996 Olympic results. Chance 9(4):28–30. https://doi.org/10.1080/09332480.1996.11884584

Sommers PM (1997) Tonga won the Atlanta Olympics! Chance 10(2):63–64. https://doi.org/10.1080/09332480.1997.10542031

StatisticsTimescom (2020) List of continents by population. https://statisticstimes.com/demographics/continents-by-population.php

Temple Lang D (2020) XML: tools for parsing and generating XML within R and S-Plus. R package version 3.99-0.3. https://CRAN.R-project.org/package=XML

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

Wang B, Chen C, Liu K (2011) The research of Olympic medal ranking methods. In: 2011 2nd international conference on artificial intelligence, management science and electronic commerce (AIMSEC). IEEE, pp 7406–7410 (in Chinese). https://doi.org/10.1109/AIMSEC.2011.6011459

Wickham H (2016) ggplot2: elegant graphics for data analysis, 2nd edn. Springer, New York

Wickham H (2019) httr: tools for working with URLs and HTTP. R package version 1.4.1. https://CRAN.R-project.org/package=httr

Wickham H, François R, Henry L, Müller K (2020) dplyr: a grammar of data manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

Wilke CO (2019) cowplot: streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.0.0. https://CRAN.R-project.org/package=cowplot

Zhao B (2017) Web scraping. In: Schintler L, McNeely C (eds) Encyclopedia of big data. Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_483-1

Zhu H (2019) kableExtra: construct complex table with 'kable' and Pipe syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra