BRAIN SCIENCE STUDY PART Nº 1

Nirupam Bidikar-1878058, Akshit Tandon-1792038, Rahul Raj Mogili-1900425

03/08/2020

Analysis 1

Construct the probability distribution function of the year of first publication of faculty.

Listing 1: Probability distribution function calculated

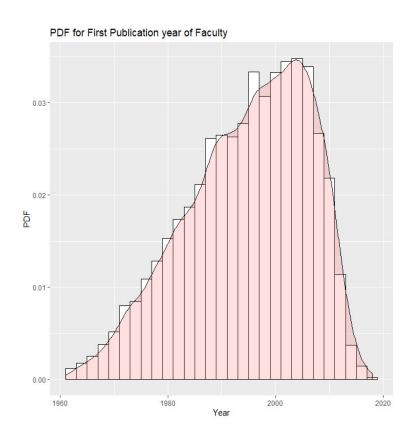


Figure 1: Probability Distribution Function of first year publications

The results indicate that we find a few outliers before the year 1960 which is statistically non relevant in this case study.

The number of publications from the above Probability Distribution Function gradually increases from the year 1960 up and until 2006, peaking at the year 2004. After the year 2006, the number of publications plummets till the time the records were updated. We also observe that the distribution is slightly skewed to the right with the later half of the distribution has more weight than the former.

The implementation of it can be found in the code snippet above. We load the author data and filter it to select minimum publishing year as 1960. We then use the histogram and density functions from ggplot to plot the PDF.

Analysis 2

Construct the probability distribution function of the total citations of faculty.

Listing 2: PDF calculated of the total citations of faculty

We are using a log operator for this analysis to help plot the Probability Distribution Function. The Probability Distribution Function is almost normally distributed. The implementation can be found in the code snippet above. We load the author data and filter it avoiding all rows with 0 citations and minimum publishing year as 1960. We use the previously used plotting functions to plot the graph.

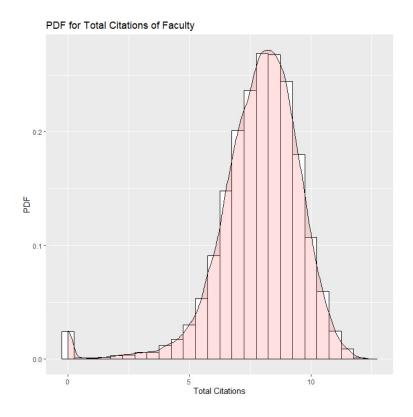


Figure 2: Probability Density Function of total citations of faculty

Analysis 3

Cluster the subject areas of faculty using the Louvain algorithm (k=5)

Listing 3: Cluster using Louvain algorithm

```
1 publication_cat <- read.csv("brain_publication_areas.csv")</pre>
2 publication_cat <- unique(publication_cat)</pre>
3
4 #forming the edge list for the graph
5 combs <- inner_join(publication_cat, publication_cat, by="eid")</pre>
6
   combs_alt <- combs %>% filter(!area.x == area.y)
7
   #creatnig a graph from the edge list
   g <- graph_from_data_frame(d = combs %>% select(area.x,area.y), directed =←
        FALSE)
10
11
   #clustering
   clusters <- cluster_louvain(g)</pre>
12
13
14 area_to_cluster_map <- cbind(V(g)$name,clusters$membership)</pre>
15 colnames(area_to_cluster_map) <- c("area", "cluster")</pre>
16 area_to_cluster_map <- as.data.frame(area_to_cluster_map)</pre>
17 allclusters <- area_to_cluster_map</pre>
```

```
18 area_to_freq_map <- as.data.frame(table(publication_cat$area))</pre>
19 colnames(area_to_freq_map) <- c("area", "freq")</pre>
20
21 # selecting top 5 clusters
22 y <- clusterdata %>% group_by(cluster) %>% add_tally(sort=TRUE) %>% select↔
       (cluster,n)
23 y <- unique(y)
24 top5 <- c(7,3,1,2,5)
25
26 merged_areas <- merge(area_to_freq_map, area_to_cluster_map, by = "area")</pre>
27 merged_areas <- merged_areas %>% filter(merged_areas$cluster %in% top5)
28 write.csv(merged_areas, "final_clusters_w_freq.csv")
29
30 # to make a word cloud -->
31 # you need to change these 3 variables for every cluster based on your \leftarrow
       personal choice
32 cluster num = 5
33 \min_{\text{freq}} = 1
34 \text{ max\_size} = 1
35 clrs = "black"
36 particular_cluster <- filter(merged_areas, cluster == cluster_num)</pre>
   wordcloud(words = particular_cluster$area,
37
              freq = particular_cluster$freq,
38
39
              scale = c(max_size, 0.5),
40
              min.freq = min_freq,
41
              max.words = Inf,
42
              random.order = FALSE,
              rot.per = .0,
43
              ordered.colors = TRUE,
44
45
46
              use.r.layout = FALSE)
```

Out of 330 topics and 11 clusters, the top 5 clusters we get are:

- 1. Neurology
- 2. Electrical and Electronic Engineering
- 3. Environmental Chemistry
- 4. Immunology
- 5. Cognitive Neuroscience

We encountered difficulty in making the graph for the clustering algorithm and we noticed that a main cluster was getting split into two separate clusters which was the biochemistry cluster. Process Chemistry and Technology
Renewable Energy, Sustainability and the Environment
Health Information Management
Safety, Risk, Reliability and Quality
Human-Computer Interaction
Geometry and Topology
Energy Engineering and Power Technology
Civil and Structural Engineering Information Systems
Energy (miscellaneous)
Computer Networks and Communications
Filtration and Separation
Computer Vision and Pattern Recognition
Hardware and Architecture Artificial Intelligence
Enuminication Physical and Theoretical Chemistry Numerical Analysis
Instrumentation
Physics and Astronomy (all)
Cocan Engineering
Industrial Relations Mechanics of Materials
Bioengineering Biomedical Engineering Electrochemistry
Fuel Technology Electronic, Optical and Magnetic Materials
Signal Processing
Chemistry (all) Inorganic Chemistry
Electrical and Electronic Engineering
Energy (all) Condensed Matter Physics Automotive Engineering
Energy (all) Condensed Matter Physics Automotive Engineering
Surfaces, Coatings and Films Engineering (all)
Mathematics (all) Mechanical Engineering Suffaces and Interfaces
Acoustics and Ultrasonics Materials Chemistry Metals and Alloys Architecture
Atomic and Molecular Physics, and Optics
Ceramics and Composites
Applied Mathematics
Computational Mechanics Control and Systems Engineering (all)
Control and Optimization Computer Science (all) Polymers and Plastics
Computational Mechanics Control and Systems Engineering Media Technology
Theoretical Computer Science Engineering (all)
Control and Optimization Teory and Mathematics
Computational Theory and Mathematics
Fuld Flow and Transfer Processes
Industrial and Manufacturing Engineering Management Information Systems
Computer Graphics and Computer-Aided Design
Colloid and Surface Chemistry Computer Science (miscellaneous)
Statistical and Nonlinear Physics Materials Science (miscellaneous)

Figure 3: Word Cloud 1

Archeology (arts and humanities)
Conservation Ecological Modeling
Nature and Landscape Conservation
Economic Geology Oceanography
Urban Studies Earth-Surface Processes
Geochemistry and Petrology
Computers in Earth Sciences Astronomy and Astrophysics
Earth and Planetary Sciences (all)
Museology Space and Planetary Science
Geophysics Pollution Soil Science
Environmental Chemistry
Forestry Ecology Geology Archeology
Environmental Engineering Paleontology
Aquatic ScienceAtmospheric Science
Water Science and Technology Classics
Waste Management and Disposal History
Geography, Planning and Development Stratigraphy
Management, Monitoring, Policy and Law
Earth and Planetary Sciences (miscellaneous)
Global and Planetary Change
Environmental Science (miscellaneous)
Tourism, Leisure and Hospitality Management

Figure 4: Word Cloud 2

Dentistry (miscellaneous)
Community and Home Care
Physical Therapy, Sports Therapy and Rehabilitation
Family Practice Otorhinolaryngology
Advanced and Specialized Nursing
Occupational Therapy Orthopedics and Sports Medicine
Histology Geriatrics and Gerontology
Podiatry Pediatrics, Perinatology and Child Health
Developmental Neuroscience Surgery Aging
Orthodontics Neuroscience (all) Periodontics
Neurology (Clinical)
Embryology Neurology Anatomy Oral Surgery
Cardiology and Cardiovascular Medicine
Developmental Biology Rehabilitation
Pathology and Forensic Medicine
Anesthesiology and Pain Medicine
Pulmonary and Respiratory Medicine
Critical Care and Intensive Care Medicine
Emergency Medicine

Figure 5: Word Cloud 3

Leadership and Management
Care Planning Nursing (all)
Transplantation
Equine Veterinary (all)
Parasitology Microbiology
Epidemiology Virology
Immunology and Allergy
Chiropractics Immunology Safety Research
Infectious DiseasesFood Animals
Public Health, Environmental and Occupational Health
Microbiology (medical) Nephrology
Animal Science and ZoologySmall Animals
Insect ScienceRheumatology
Health Policy
Veterinary (miscellaneous)
Fundamentals and Skills

Figure 6: Word Cloud 4

Economics, Econometrics and Finance (miscellaneous)
Political Science and International Relations
Business, Management and Accounting (all)
Social Sciences (miscellaneous)
Arts and Humanilles (all)
Finance Life-span and Life-course Studies
Development
Anthropology
Health Professions (all) Health (social science)
Literature and Literary Theory Social Sciences (all) Education Music
Speech and Hearing Language and Linguistics
Decision Sciences (all) — Psychology (all) Philosophy
Developmental and Educational Psychology
LPN and LVN Behavioral Neuroscience Public Administration
Cognitive Neuroscience
Experimental and Cognitive Psychology
Neuropsychology and Physiological Psychology
Arts and Humanities (miscellaneous) Demography
Linguistics and Language
Marketing Social Psychology
Accounting Sociology and Political Science Religious Studies
Psychology (miscellaneous) Cultural Studies
Economics and Econometrics Gender Studies
Issues, Ethics and Legal Aspects
Business and International Management
Economics, Econometrics and Finance (all)
Business, Management and Accounting (miscellaneous)
Visual Arts and Performing Arts
Organizational Behavior and Human Resource Management

Figure 7: Word Cloud 5

Analysis 4

Generate a bar plot of the number of publications per subject area per continent pre-2014, post-2014, and their difference.

Listing 4: Generating bar plot of the number of publications per subject

```
1 clusterdata <- read.csv("final_clusters_w_freq.csv")</pre>
 2 author <- read.csv("brain_author.csv")</pre>
 3 pub_details <- read.csv("brain_publication_details.csv")</pre>
 4 publication_cat <- read.csv("brain_publication_areas.csv")</pre>
 5
 6 combined_frame_1 <- merge(author, pub_details, by="scopus_id")</pre>
 7 combined frame 2 <- merge(combined frame 1, publication cat, by.x="eids",\leftrightarrow
       by.y="eid")
 8 mergedData <- merge(combined_frame_2, clusterdata, by.x = "area",by.y ="←
       area")
9 clusterHead <- sqldf("select area,cluster,max(Freq)from clusterdata group ←
       by cluster")
10 clusterHead
11 finalDataset <- merge(mergedData,clusterHead,by="cluster")</pre>
12 #this is a 2.6 GB file
13 write.csv(finalDataset,"finalset.csv")
14
15 # Final Dataset to use after all joins
16
17 finalClusterSet <- read.csv("finalset.csv")</pre>
18 finalClusterSet <- finalClusterSet %>% select(eids,area.y,cip_title,←
       scopus_id,pub_year,region,)
19
20 after_2014_q4 <- filter(finalClusterSet, pub_year > 2014)
21 before_2014_q4 <- filter(finalClusterSet, pub_year <= 2014)
22
23
24 #filtering values before and after 2014 and grouping based on subject area↔
        and region
25 values_after_2014 <- after_2014_q4 %>% group_by(after_2014_q4$area.y, ←
       region) %>% filter(!is.na(region)) %>% add_tally()
26
27 values_before_2014 <- before_2014_q4 %>% group_by(before_2014_q4\alpha, \leftarrow
       region) %>% filter(!is.na(region)) %>% add_tally()
28
29
30 #plotting after 2014
31 g1 <- ggplot(values_after_2014, aes(x = region, y = n, fill =\leftarrow
       values_after_2014$area.y ))+
```

```
32
      geom_bar(stat = "identity",position=position_dodge())+ coord_flip() + ←
         ggtitle("Number of publications per subject area per continent post←
          -2014") +
      labs(x = "region", y= "publications", fill="Subject Area")
33
34
35
   #plotting beofre 2014
36
   g2 <- ggplot(values_before_2014, aes(x = region, y = n, fill =←
      values_before_2014$area.y ))+
      geom_bar(stat = "identity",position=position_dodge())+ coord_flip() + ←
37
          ggtitle("Number of publications per subject area per continent pre↔
          -2014") +
      labs(x = "region", y= "publications", fill="Subject Area")
38
39
40
41
   grid.arrange(g1,g2)
```

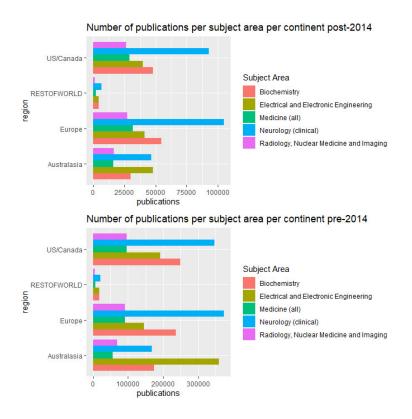


Figure 8: Number of Publications per continent pre and post 2014

- 1. Neurology has most publications in all the regions except for Australia in both pre-2014 and post-2014.
- 2. Meanwhile, Australasia has most number of publication in the subject area of Electrical and Electronic Engineering than any other region in both pre and post 2014
- 3. Europe has the most number of publications for a given subject area which is Neurology compared to any other continent.

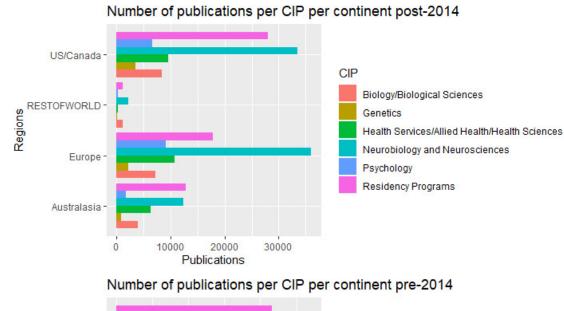
4. The Rest of the World too has Neurology being the most number of publications.

Analysis 5

Generate a bar plot of the number of publications per CIP category per continent pre-2014, post-2014, and their difference.

Listing 5: Generating bar plot of the number of publications per CIP category per continent

```
1 author_data <- read.csv("brain_author.csv")</pre>
2 pub_details <- read.csv("brain_publication_details.csv")</pre>
4 combined_frame <- merge(author_data, pub_details, by="scopus_id")</pre>
5 freq_cip <- as.data.frame(table(combined_frame$cip_title))</pre>
6 colnames(freq_cip) <- c("Cip_Title","Freq")</pre>
7 freq_cip_top6 <- sqldf("select * from freq_cip order by Freq Desc limit 6"←
       )
8 dd <- merge(combined_frame,freq_cip_top6 , by.x = "cip_title",by.y = "←
       Cip Title")
9 new_dd <- sqldf("select * from dd where region <> 'NA' ")
10
11 after_2014 <- filter(new_dd, pub_year > 2014) %>% filter(!is.na(region)) ←
       %>% filter(!is.na(cip_title))
12 grouped_after_2014 <- after_2014%>% group_by(after_2014$region, ←
       after_2014$cip_title) %>% add_tally()
13
14 before_2014 <- filter(new_dd, pub_year < 2014) %>% filter(!is.na(region)) ←
        %>% filter(!is.na(cip_title))
15 grouped_before_2014 <- before_2014%>% group_by(before_2014$region, ←
       before_2014$cip_title) %>% add_tally()
16
17 comb_data <- merge(grouped_after_2014,grouped_before_2014, by.x="cip_title↔
       " , by.y="region")
18
19
20 # plot for after 2014
21 g1 <- ggplot(grouped_after_2014, aes(x = region, y = n, fill = cip_title))\leftarrow
     geom_bar(stat = "identity",position=position_dodge())+ coord_flip() +
22
       ggtitle("Number of publications per CIP per continent post-2014") + ←
23
           labs(x = "Region", y = "Publications", fill ="CIP")
24
25 #plot for before 2014
26 g2 <- ggplot(grouped_before_2014, aes(x = region, y = n, fill = cip_title)\leftarrow
       )+
```



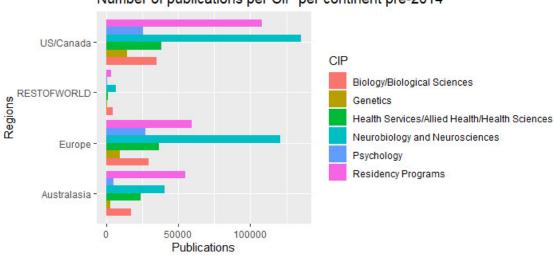


Figure 9: Number of Publications Pre and Post 2014

- 1. The sum of all the CIP publications in US/Canada is greater than any other continent since the year 1960.
- 2. Neurobiology and Neurosciences has the highest number of publications not only in Europe but all over the world pre-2014 whereas US/Canada takes the lead post-2014 in the same field overall.