Report

Martina Le-Bert Heyl

2022-04-04

Shiny url https://mlebertheyl.shinyapps.io/stat294_H2/

My Github https://github.com/mlebertheyl

Github repository for this project https://github.com/mlebertheyl/stat294_H2

Motivation

Multiple studies oriented to anthropogenic threats have attempted to systematize the scientific evidence generated in Chile on the impact of human activities on ecosystems. Despite efforts, these focus on specific problems and fail to classify threats according to their greater or lesser relevance among the different ecosystems. This presents an opportunity to evaluate the available information in a quantitative and systematic way in, so it reflects research priorities in the area. Moreover, evaluating these priorities, and what scientists are focused on, should give us a better understanding of the underlying anthropogenic threats that have affected Chilean ecosystems through out the years. Bibliometrics is a set of qualitative and quantitative procedures used to analyze the academic literature and, ultimately, find generalized patterns and dynamics in a given research discipline.

Purpose of the Shiny app.

- Collect all the relevant available information on the topic of interest, and evaluate the publication trend over the years.
- Evaluate the publication trend by ecosystem, and type of threat.
- Evaluate the most used keywords by decade in published articles.
- Present the information in an appealing and simple manner.

Future goals

- Adding bibliometric indicators: main publication journals; main author's affiliation; research productivity; word bigrams; and word co-occurrence network in abstracts.
- Allow the Shiny app to upload data from WoS for any region or country.
- Allowing for more than one data set to be upload, so regions or countries can be compared.

Brief description on how the data was produced.

Search

The literature search was done through the ISI Web of Science (WoS) main collection database (https://www.webofknowledge.com), using the a Boolean search; `TS= ((human OR anthropogenic OR mining OR industrial* OR agricult* OR domestic) AND (pollut* OR contaminant* OR threat* OR disturbanc* ) AND (chile*))`. From this we obtained a BibTeX file including: author(s); author affiliation(s); journal; publication title; keywords; abstract; year of publication; and citation count.

Article selection

Exclude articles that are not related to the topic of interest based on: title (1); abstract (2); research results (3).

Bibliometric Analysis

Relevant bibliometric indicators obtained with the `R` package `Bibliometrix` were: publications per year, and author keywords.

Content Analysis and Topic Identification

To identify and group the documents into different topics an LDA (Latent Dirichlet Assignment) model was used, together with manual classification.

Brief description on how the app was built.

Packages

```
library(readxl)
library(tidyverse)
library(shiny)
library(plotly)
library(bibliometrix)
library(DT)
library(shinydashboard)
library(quanteda)
library(topicmodels)
library(tools)
library(quanteda.textplots)
```

Data for the app

```
data <- read_excel("data/data.xlsx")

main_information <- data.frame(Description=c("Timespan", "Documents", "Ecosystems", "Threats"),
                               Counts=c("1990-2021","760", "6","6"))

df1<-data %>%
  filter(!is.na(Ecosystem))%>%
  group_by(Ecosystem) %>%
  count(PY) %>%
  mutate(cumulative = cumsum(n))

df2<-data %>%
```

```
  filter(!is.na(Threat))%>%
  group_by(Threat) %>%
  count(PY) %>%
  mutate(cumulative = cumsum(n))

conteo<-data%>%
  count(PY)%>%
  mutate(cumulative=cumsum(n)) %>%
  mutate(percentage=100*(n / sum(n)))%>%
  mutate(percentage_cum=100*(cumulative / sum(n))) %>%
  mutate(across(where(is.numeric), round, 2))


## Data Keywords
data_1999<-data %>%
  filter(!is.na(AB))%>%
  filter(PY< 2000)

data_2009<-data%>%
  filter(!is.na(AB))%>%
  filter(PY>1999 & PY< 2010)

data_2021<-data %>%
  filter(!is.na(AB))%>%
  filter(PY> 2009)

corpus_abstract_1999 <- corpus(data_1999$AB, docnames = data_1999$doc_id, docvars = data.frame(year = data
_1999$PY))
corpus_abstract_2009 <- corpus(data_2009$AB, docnames = data_2009$doc_id, docvars = data.frame(year = data
_2009$PY))
corpus_abstract_2021 <- corpus(data_2021$AB, docnames = data_2021$doc_id, docvars = data.frame(year = data
_2021$PY))

list_stopwords <- readLines("list_stopwords.csv", encoding = "UTF-8")
other_stopwords<- c("elsevier","b.v","rights"," reserved","-1","-2","yr")
original_word<-c("areas","soils","rivers","activities","concentration","level","cu","sources","risks","eff
ects","sample","pb","higher","value","zn","forest","mma","dma")
new_word<-c("area","soil","river","activity","concentrations","levels","copper","source","risk","effect","
samples","lead","high","values","zinc","forests","monomethylarsonic acid","dimethylarsinic acid")

toks_1999 <- corpus_abstract_1999 %>%
  tokens(remove_punct = TRUE,remove_numbers = TRUE, remove_symbols = TRUE) %>%
  tokens_tolower() %>%
  tokens_remove(stopwords("english")) %>%
  tokens_select(c(-1,list_stopwords,other_stopwords), selection = "remove", padding = FALSE) %>%
  tokens_replace(pattern=original_word,
                 replacement=new_word)

toks_2009 <- corpus_abstract_2009 %>%
  tokens(remove_punct = TRUE,remove_numbers = TRUE, remove_symbols = TRUE) %>%
  tokens_tolower() %>%
  tokens_remove(stopwords("english")) %>%
  tokens_select(c(-1,list_stopwords,other_stopwords), selection = "remove", padding = FALSE) %>%
  tokens_replace(pattern=original_word,
                 replacement=new_word)

toks_2021 <- corpus_abstract_2021 %>%
  tokens(remove_punct = TRUE,remove_numbers = TRUE, remove_symbols = TRUE) %>%
  tokens_tolower() %>%
  tokens_remove(stopwords("english")) %>%
  tokens_select(c(-1,list_stopwords,other_stopwords), selection = "remove", padding = FALSE) %>%
  tokens_replace(pattern=original_word,
                 replacement=new_word)


wfreq_1999 <- toks_1999 %>%
  unlist()%>%
  table() %>%
  as.data.frame() %>%
```

```
  arrange(desc(Freq)) %>%
  head(20) %>%
  rename("unigram 1990-1999" = 1,frequency = 2)

wfreq_2009 <- toks_2009 %>%
  unlist()%>%
  table() %>%
  as.data.frame() %>%
  arrange(desc(Freq)) %>%
  head(20) %>%
  rename("unigram 2000-2009" = 1,frequency = 2)

wfreq_2021 <- toks_2021 %>%
  unlist()%>%
  table() %>%
  as.data.frame() %>%
  arrange(desc(Freq)) %>%
  head(20) %>%
  rename("unigram 2010-2021" = 1,frequency = 2)
```

ui and server

```
ui <- dashboardPage(
  skin = "red",
  dashboardHeader(title = "Literature Review"),

  dashboardSidebar(sidebarMenu(

    menuItem("Database", tabName = "database", icon = icon("database"), badgeLabel = "readme", badgeColor
= "green"),
    menuItem("Main Information", tabName = "information", icon = icon("database")),
    menuItem("Ecosystems", tabName = "ecosystem", icon = icon("dashboard")),
    menuItem("Threats", tabName = "threat", icon = icon("dashboard")),
    menuItem("Keywords", tabName = "keywords", icon = icon("dashboard")))),

  dashboardBody(

    tabItems(
      tabItem(tabName = "database",
              h2("Anthropogenic threats on Chilean Ecosystems"),
              fluidRow(
                infoBox(width = 12,"Literature Review", "Anthropogenic threats to Chilean ecosystems: a sy
stematic literature review", icon = icon("list"), color = "red", fill = F),
                infoBox(width = 12,"Web Of Science", "1990-2021", icon = icon("list"), color = "red", fill
= TRUE),
                infoBox(width = 12,"boolean search", "TS= ((human OR anthropogenic OR mining OR industrial
* OR agricult* OR domestic)  AND  (pollut* OR contaminat OR threat* OR disturbanc* )  AND (chile*))",
                        icon = icon("list"), color = "orange", fill = TRUE),
                infoBox(width = 12,"Research Articles", "760", icon = icon("list"), color = "yellow", fill
= TRUE),
                infoBox(width = 12,"Manual classification", "Ecosystems and Threats", icon = icon("list"),
color = "teal", fill = TRUE),
                infoBox(width = 12,"Temporal Trends", "Number of publications", icon = icon("list"), color
= "aqua", fill = TRUE),
                infoBox(width = 12,"Text Mining", "Most frequent words", icon = icon("list"), color = "lig
ht-blue", fill = TRUE))),

      tabItem(tabName = "information",
              h2("Main information from publications included in the database."),
              fluidRow(
                box(width = 4, DTOutput(outputId = "table"),status = "warning", solidHeader = TRUE,
                    collapsible = TRUE)),
              h2("Annual number of publications from 1990 to 2021."),
              fluidRow(
                box(plotlyOutput(outputId = "conteo"),status = "info", solidHeader = TRUE,
                    collapsible = TRUE))),

      tabItem(tabName = "ecosystem",
```

```
              h2("Cumulative number of publications per Ecosystem"),
              fluidRow(
                box(width = 4, radioButtons("Ecosystem", "Ecosystem:",
                                            c("Terrestrial", "Urban", "Freshwater","Marine","Glaciers", "M
ultiple")),status = "info", solidHeader = TRUE,
                    collapsible = TRUE),
                box(plotlyOutput(outputId = "plot1"),status = "info", solidHeader = TRUE,
                    collapsible = TRUE))),

      tabItem(tabName = "threat",
              h2("Cumulative number of publications per Antrhopogenic Threat"),
              fluidRow(
                box(width = 4, radioButtons("Threat", "Threat",
                                            c("Pollution", "Habitat change", "Overexploitation","Invasive
alien species","Climate change", "Multiple")),
                    status = "info", solidHeader = TRUE,
                    collapsible = TRUE),
                box(plotlyOutput(outputId = "plot2"),status = "info", solidHeader = TRUE,
                    collapsible = TRUE))),

      tabItem(tabName = "keywords",
              h2("Top 20 most relevant words based on publishes abstracts"),
              fluidRow(
                box(width=8,title = "1990-2000", status = "primary", solidHeader = TRUE,
                    collapsible = TRUE, collapsed = TRUE,
                    plotlyOutput(outputId = "plot3")),
                box(width=8,title = "2001-2009", status = "primary", solidHeader = TRUE,
                    collapsible = TRUE, collapsed = TRUE,
                    plotlyOutput(outputId = "plot4")),
                box(width=8,title = "2010-2021", status = "primary", solidHeader = TRUE,
                    collapsible = TRUE, collapsed = TRUE,
                    plotlyOutput(outputId = "plot5"))
              )
      )
    )))


server <- function(input, output) {

  output$conteo<-renderPlotly({
    ggplotly({
      plot_conteo<-ggplot(conteo, aes(x = PY, y = n))+
        geom_line(color="#00BFFF", size=0.5,alpha =0.5)+
        geom_point(aes(x = PY, y = n), color="#00BFFF",size =2)+
        scale_x_continuous(breaks=c(1990,1995,2000,2005,2010,2015, 2021),limits=c(1989, 2021))+
        scale_y_continuous(breaks=c(0,30,60,90,120),limits=c(0,120))+
        labs(x="", y="Number of publications")+ xlab("Year")+
        theme(axis.text =element_text(size=12,color="#191919"),
              axis.title=element_text(size=12,color="#191919"),
              panel.background = element_blank(),
              panel.border = element_rect(color="#191919", size=1, fill=NA),
              legend.position = "none")
      plot_conteo

    })
  })

  filtered_data1 <- reactive({
    subset(df1,
           Ecosystem %in% input$Ecosystem)})

  output$plot1 <- renderPlotly({
    ggplotly({
      p1 <- ggplot(filtered_data1(), aes(x = PY, y = cumulative))+
        geom_line(color="#00BFFF", size=0.5,alpha =0.5)+
        geom_point(aes(x = PY, y = cumulative), color="#00BFFF",size =2)+
        scale_x_continuous(breaks=c(1990,1995,2000,2005,2010,2015, 2021),limits=c(1989, 2021))+
        scale_y_continuous(breaks=c(0,30,60,90,120,150,180,210,240),limits=c(0,240))+
        labs(title=NULL,x="", y="Cumulative number of publications")+ xlab("Year")+
```

```
      theme(axis.text =element_text(size=12,color="#191919"),
            axis.title=element_text(size=12,color="#191919"),
            panel.background = element_blank(),
            panel.border = element_rect(color="#191919", size=1, fill=NA),
            legend.position = "none")

    p1

  })
})

filtered_data2 <- reactive({
  subset(df2,
         Threat %in% input$Threat)})

output$plot2 <- renderPlotly({
  ggplotly({
    p2 <- ggplot(filtered_data2(), aes(x = PY, y = cumulative))+
      geom_line(color="#00BFFF", size=0.5,alpha =0.5)+
      geom_point(aes(x = PY, y = cumulative), color="#00BFFF",size =2)+
      scale_x_continuous(breaks=c(1990,1995,2000,2005,2010,2015, 2021),limits=c(1989, 2021))+
      scale_y_continuous(breaks=c(0,40,80,120,160,200,240,280,320,360,400),limits=c(0,400))+
      labs(title=NULL,x="", y="Cumulative number of publications")+ xlab("Year")+
      theme(axis.text =element_text(size=12,color="#191919"),
            axis.title=element_text(size=14,color="#191919"),
            panel.background = element_blank(),
            panel.border = element_rect(color="#191919", size=1, fill=NA),
            legend.position = "none")

    p2
  })
})
output$table <- renderDT(main_information,
                         options = list(dom = 't'))

output$plot3 <- renderPlotly({
  ggplotly({
    hist_1999<-wfreq_1999 %>%
      rename(unigram="unigram 1990-1999") %>%
      ggplot(aes(x = reorder(unigram, -frequency, mean), y = frequency)) +
      geom_bar(stat = "identity") +
      labs(x=NULL, y="Counts")+
      theme(axis.text =element_text(size=10,color="#191919"),
            axis.text.x = element_text(angle = 45,hjust = 1),
            axis.title=element_text(size=12,color="#191919"),
            plot.title = element_text(color="#191919", hjust=0),
            panel.background = element_blank())
  })
})

output$plot4 <- renderPlotly({
  ggplotly({
    hist_2009<-wfreq_2009 %>%
      rename(unigram="unigram 2000-2009") %>%
      ggplot(aes(x = reorder(unigram, -frequency, mean), y = frequency)) +
      geom_bar(stat = "identity") +
      labs(x=NULL, y="Counts") +
      theme(axis.text =element_text(size=10,color="#191919"),
            axis.text.x = element_text(angle = 45,hjust = 1),
            axis.title=element_text(size=12,color="#191919"),
            plot.title = element_text(color="#191919", hjust=0),
            panel.background = element_blank())
  })
})

output$plot5 <- renderPlotly({
  ggplotly({
    hist_2021<-wfreq_2021 %>%
      rename(unigram="unigram 2010-2021") %>%
```

```
        ggplot(aes(x = reorder(unigram, -frequency, mean), y = frequency)) +
        geom_bar(stat = "identity") +
        labs(x=NULL, y="Counts")+
        theme(axis.text =element_text(size=10,color="#191919"),
              axis.text.x = element_text(angle = 45,hjust = 1),
              axis.title=element_text(size=12,color="#191919"),
              plot.title = element_text(color="#191919", hjust=0),
              panel.background = element_blank())
    })
  })

}

shinyApp(ui = ui, server = server)
```

Results

How to use the app

- Sidebar with bibliometric indicators.
- Cumulative number of publications per ecosystem, where the ecosystem can be selected by clicking on it. Ecosystem option box can be collapsed click on the upper right box corner.
- Cumulative number of publications per threat, where the anthropogenic source of threat can be selected by clicking on it. The option box can be collapsed click on the upper right box corner.
- Histogram with most important keywords by decade is displayed by clicking on the upper right box corner of the desired time span.
- Moving the cursor over the figures shows relevant information for each observation point.

Results interpretation, and conclusion

The chosen bibliometric indicators give an overview of the general research trends throughout the years. It also possible to see in the figures which specific topics have been more *attractive* to researchers in different years, and if these have continued to be a topic of research or not.

The goal of the app was to make a tool in which could be easily seen how publication trends related to impact of human activities on ecosystems has evolved. Consequently, statistical analyses over these trends were not made, so result interpretation is open to the user of the app. However, one would like to think that scientific research is mainly driven by the necessity of studying new dynamics and processes. In a similar way, one could think, scientists are the first to be interest in new, and arising processes, because there's no other purpose than to expand knowledge. Under this paradigm, results could be interpreted as environmental concerns a country or region is facing, and how this *concern* might be evolving through out the years.

The app built focused on Chile, meaning the ecosystems and threats won't be the same for other countries. It is somewhat clear, that an important ecosystem in this country are glaciar ecosystems, however one wouldn't imagine this to be also true for tropical regions. In line with this, it is to be interpreted that the most relevant ecosystems in Chile are terrestrial, urban, freshwater, and marine ecosystems. The `multiple` option in the `Ecosystem` box, refers to studies including more than one ecosystem. A similar interpretation could be applied for threats.

Nevertheless, it is important to know more bibliometric indicators to gain a better insight. For small countries like Chile, a better interpretation, for example, could be made if the author's affiliation was considered. One could imagine, that even though some topic doesn't have the majority of research associated to it, maybe the affiliation or university behind the research is almost always the same. In Chile, there's more or less one university per region - except for the main city-, so if an affiliation is recurrent on one particular topic, we could interpret it as this topic being important in the region where the university is. This, and other examples, explain what it would de important to add more bibliometric indicators. Finally, I will add, that this indicators were actually calculated, but omitted for this project, not to over extend the project and presentation for the course.