Network Citation Analysis

R code

Arnald Puy

Contents

1 Preliminary

```
# PRELIMINARY FUNCTIONS ########################################################

sensobol::load_packages(c("sensobol", "data.table", "tidyverse", "janitor",
                          "igraph", "ggraph", "tidygraph", "cowplot", "viridis",
                          "wesanderson", "parallel", "doParallel", "tm"))

# Custom theme for plots
theme_AP <- function() {
  theme_bw() +
    theme(panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.background = element_rect(fill = "transparent",
                                           color = NA),
          legend.key = element_rect(fill = "transparent",
                                    color = NA),
          strip.background = element_rect(fill = "white"),
          legend.margin = margin(0.5, 0.1, 0.1, 0.1),
          legend.box.margin = margin(0.2,-4,-7,-7),
          plot.margin = margin(3, 4, 0, 4),
          legend.text = element_text(size = 8),
          axis.title = element_text(size = 10),
          legend.key.width = unit(0.4, "cm"),
          legend.key.height = unit(0.4, "cm"),
          legend.title = element_text(size = 9))
}

# Define color palette
selected_wesanderson <- "Chevalier1"
```

```
water.models <- c("WaterGAP", "PCR-GLOBWB", "LPJmL", "CLM4.5", "DBHM",
                  "TOPMODEL", "H08", "JULES-W1", "MPI-HM", "VIC", "SWAT",
                  "GR4J", "HYPE", "HBV", "MATSIRO", "SACRAMENTO", "MHM",
                  "CWatM", "ORCHIDEE")

dt <- list()

for (i in 1:length(water.models)) {

  dt[[i]] <- fread(paste(water.models[[i]], ".csv", sep = ""), skip = 1) %>%
    clean_names() %>%
    data.table()

}

names(dt) <- water.models
dt.water <- rbindlist(dt, idcol = "Model")
```

```r
wos.dt <- fread("final.dt.csv")
wos.titles <- wos.dt[Model %in% water.models]

# REMOVE DUPLICATED REFERENCES ##################################################

# Number of papers in more than one model
n_occur <- data.frame(table(dt.water$publication_id))
papers_repeated <- data.table(n_occur[n_occur$Freq > 1,])
length(papers_repeated$Var1) # number of repeated papers
```

```
## [1] 2323
```

```r
# Fraction of repeated papers over the total
length(papers_repeated$Var1) / nrow(dt.water)
```

```
## [1] 0.07791903
```

```r
# How many papers are repeated twice, three times, etc...
papers_repeated[, .(N.repeated.papers = .N), Freq]
```

```
##     Freq N.repeated.papers
##    <int>             <int>
## 1:     2              1798
## 2:     4               106
## 3:     6                18
## 4:     3               348
## 5:     5                38
## 6:     8                 5
## 7:     7                 6
## 8:     9                 1
## 9:    11                 3
```

```r
# Extract which papers are repeated for which model
dt.sample.repeated <- dt.water[publication_id %in% papers_repeated$Var1] %>%
  .[, .(publication_id, Model, title, source_title_anthology_title)] %>%
  .[order(publication_id)]

dt.sample.repeated
```

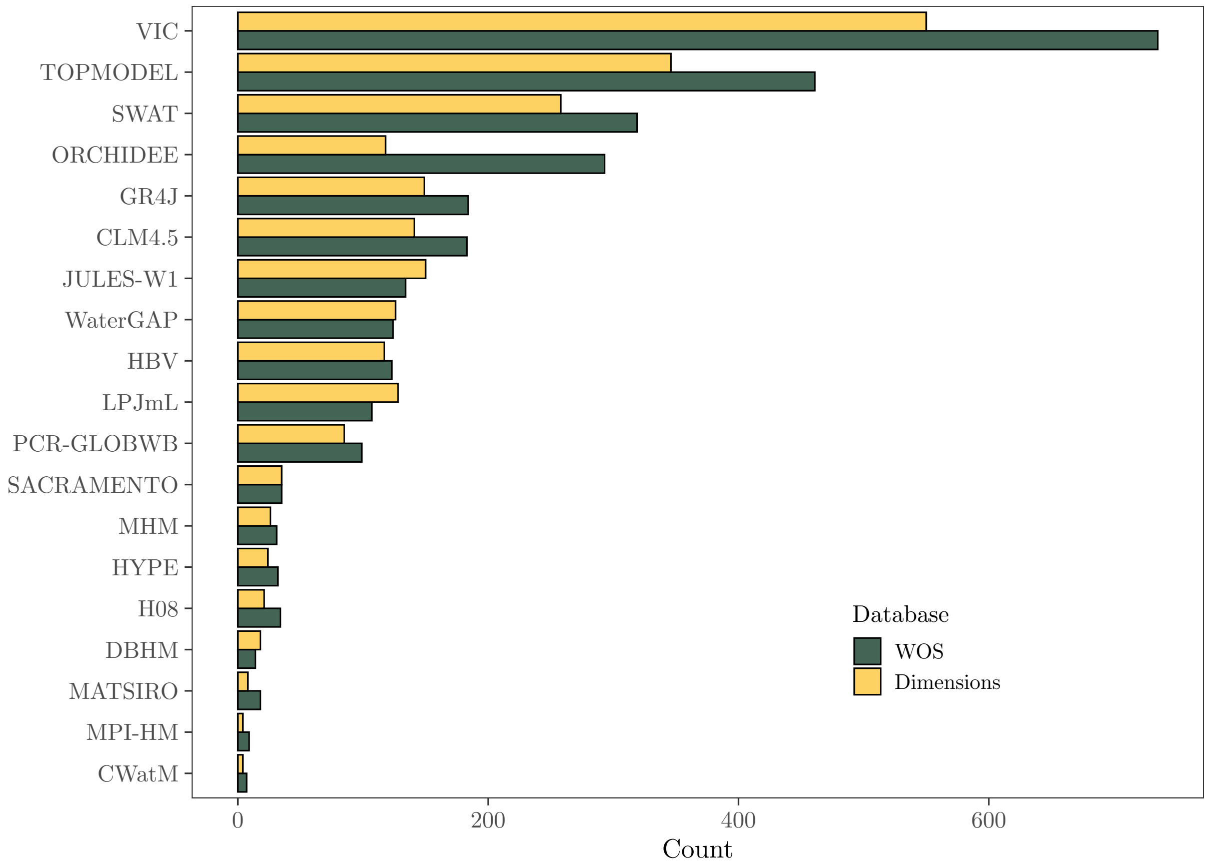
```
##       publication_id      Model
##               <char>     <char>
##    1: pub.1000120678   TOPMODEL
##    2: pub.1000120678 SACRAMENTO
##    3: pub.1000226548   WaterGAP
##    4: pub.1000226548 PCR-GLOBWB
##    5: pub.1000226548        HBV
##   ---
## 5482: pub.1167654662 PCR-GLOBWB
## 5483: pub.1167736853 PCR-GLOBWB
## 5484: pub.1167736853        MHM
## 5485: pub.1167835489     CLM4.5
## 5486: pub.1167835489   TOPMODEL
##
##
##    1:                                        Temporal dynamics of model parameter sensitivi
##    2:                                        Temporal dynamics of model parameter sensitivi
##    3:                                                                             Multiscale
##    4:                                                                             Multiscale
##    5:                                                                             Multiscale
##   ---
## 5482: Scenario setup and forcing data for impact model evaluation and impact attribution wi
## 5483:         Tradeoffs Between Temporal and Spatial Pattern Calibration and Their Impacts o
## 5484:         Tradeoffs Between Temporal and Spatial Pattern Calibration and Their Impacts o
## 5485:                                                                Development of inter-g
## 5486:                                                                Development of inter-g
##          source_title_anthology_title
##                                <char>
##    1:        Water Resources Research
##    2:        Water Resources Research
##    3:     Journal of Hydrometeorology
##    4:     Journal of Hydrometeorology
##    5:     Journal of Hydrometeorology
##   ---
## 5482: Geoscientific Model Development
## 5483:        Water Resources Research
## 5484:        Water Resources Research
## 5485: Geoscientific Model Development
## 5486: Geoscientific Model Development
```

```r
# Randomly retrieve only one of the repeated studies per model
set.seed(6)
dt.no.repeated <- dt.sample.repeated[,.SD[sample(.N, min(1,.N))], publication_id]

# Setkey to filter and retrieve
res <- setkey(dt.water, publication_id, Model) %>%
  .[J(dt.no.repeated$publication_id, dt.no.repeated$Model)]

# Make the dataset without repeated papers across models
final.dt <- rbind(res, dt.water[!publication_id %in% papers_repeated$Var1])

# Check which papers do not have cited bibliography metadata and exclude them
final.dt <- final.dt[, empty_cited_references:= grepl("^\\s*$", cited_references)] %>%
  .[empty_cited_references == FALSE] %>%
  # Filter dataset to ensure all titles use a water model
  .[tolower(.$title) %in% wos.titles$title.large]

# Check the WOS and the Dimensions dataset
```

```
wos.dimensions <- merge(wos.dt[Model %in% water.models] %>%
  .[, .(WOS = .N), Model],
  final.dt[, .(Dimensions = .N), Model],
  by = "Model")

wos.dimensions[order(-Dimensions)]
```

```
##          Model   WOS Dimensions
##         <char> <int>      <int>
##  1:        VIC   735        550
##  2:   TOPMODEL   461        346
##  3:       SWAT   319        258
##  4:   JULES-W1   134        150
##  5:       GR4J   184        149
##  6:     CLM4.5   183        141
##  7:      LPJmL   107        128
##  8:   WaterGAP   124        126
##  9:   ORCHIDEE   293        118
## 10:        HBV   123        117
## 11: PCR-GLOBWB    99         85
## 12: SACRAMENTO    35         35
## 13:        MHM    31         26
## 14:       HYPE    32         24
## 15:        H08    34         21
## 16:       DBHM    14         18
## 17:    MATSIRO    18          8
## 18:      CWatM     7          4
## 19:     MPI-HM     9          4
```

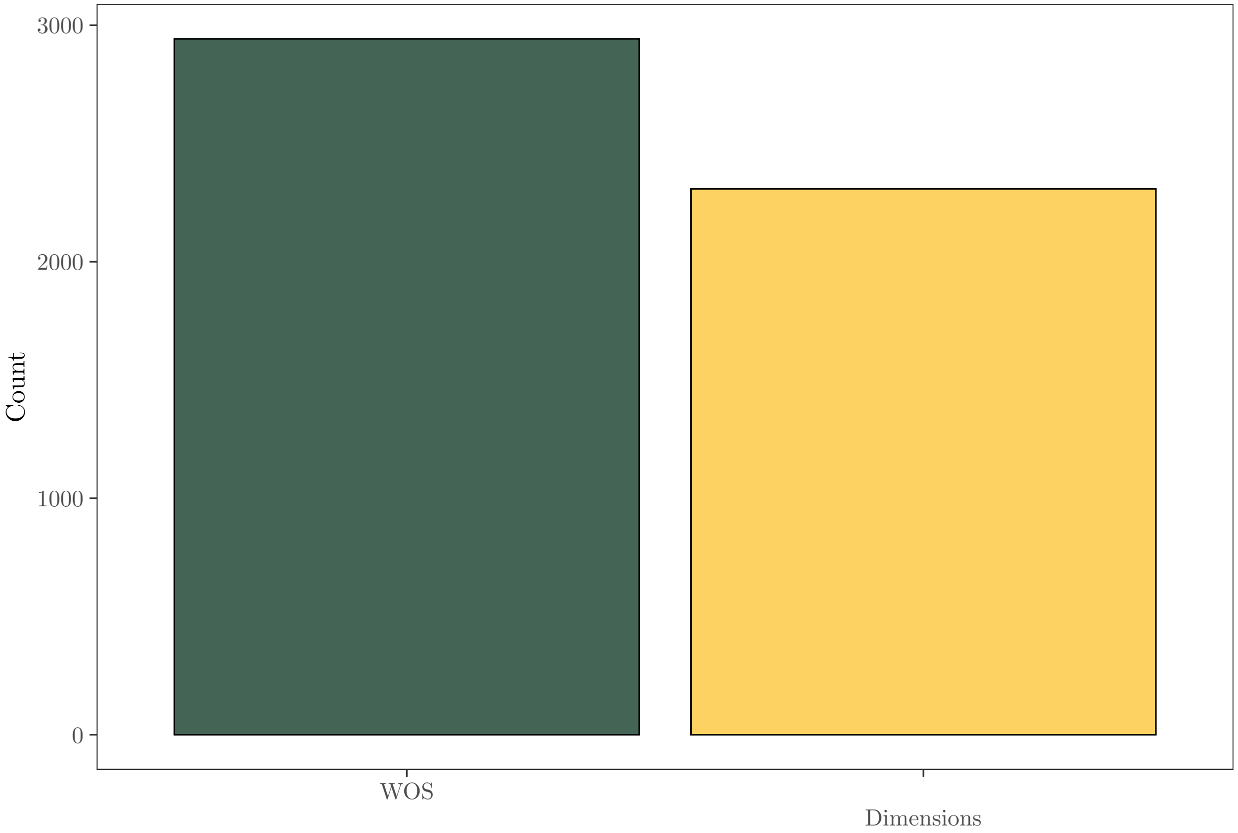
```
# PLOT DIFFERENCES BETWEEN WOS AND DIMENSIONS ##################################

plot.models <- wos.dimensions %>%
  melt(., measure.vars = c("WOS", "Dimensions")) %>%
  ggplot(., aes(reorder(Model, value), value, fill = variable)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  coord_flip() +
  scale_fill_manual(name = "Database",
                    values = wes_palette(name = selected_wesanderson, 2)) +
  labs(y = "Count", x = "") +
  theme_AP() +
  theme(legend.position = c(0.73, 0.2))

plot.models
```



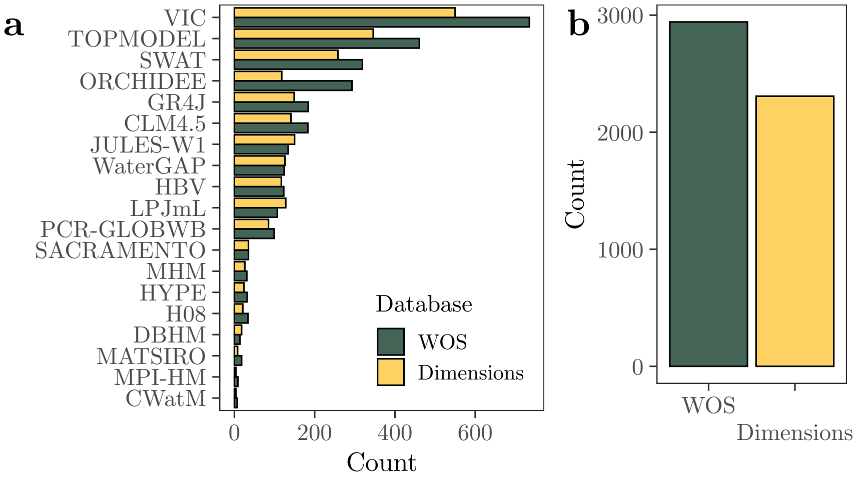
```r
wos.dimensions.dt <- wos.dimensions %>%
  melt(., measure.vars = c("WOS", "Dimensions"), variable.name = "dataset") %>%
  .[, .(total = sum(value)), dataset]

wos.dimensions.dt
```

```
##       dataset total
##        <fctr> <int>
## 1:        WOS  2942
## 2: Dimensions  2308
```

```r
plot.databases <- wos.dimensions.dt %>%
  ggplot(., aes(dataset, total, fill = dataset)) +
  geom_bar(stat = "identity", color = "black") +
  scale_fill_manual(name = "Database",
                    values = wes_palette(name = selected_wesanderson, 2)) +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  labs(x = "", y = "Count") +
  theme_AP() +
  theme(legend.position = "none")

plot.databases
```



```r
# MERGE AND PLOT ###############################################################

plot_grid(plot.models, plot.databases, ncol = 2, rel_widths = c(0.65, 0.35),
          labels = "auto")
```


a
VIC
TOPMODEL
SWAT
ORCHIDEE
GR4J
CLM4.5
JULES-W1
WaterGAP
HBV
LPJmL
PCR-GLOBWB
SACRAMENTO
MHM
HYPE
H08
DBHM
MATSIRO
MPI-HM
CWatM
Count
Database
WOS
Dimensions
b
Count
WOS
Dimensions