# Métodos de Gestión de la Información

Minería de reglas de asociación

Ángel Ríos San Nicolás

23 de agosto de 2021

## Datos y preprocesamiento

Los datos que analizaremos contienen 199523 instancias ponderadas con 41 variables de tipo demográfico y laboral del censo de EEUU tomados en los años 1994 y 1995, que están disponibles en UCI Machine Learning Repository Census Income (KDD) Data Set[1]. Los valores de los atributos son los siguientes:

```
age: continuous.
class of worker: Not in universe, Federal government, Local government, Never worked, Private,
     Self-employed-incorporated, Self-employed-not incorporated, State government, Without pay.
detailed industry recode: 0, 40, 44, 2, 43, 47, 48, 1, 11, 19, 24, 25, 32, 33, 34, 35, 36, 37, 38,
     39, 4, 42, 45, 5, 15, 16, 22, 29, 31, 50, 14, 17, 18, 28, 3, 30, 41, 46, 51, 12, 13, 21, 23,
     26, 6, 7, 9, 49, 27, 8, 10, 20.
detailed occupation recode: 0, 12, 31, 44, 19, 32, 10, 23, 26, 28, 29, 42, 40, 34, 14, 36, 38, 2,
     20, 25, 37, 41, 27, 24, 30, 43, 33, 16, 45, 17, 35, 22, 18, 39, 3, 15, 13, 46, 8, 21, 9, 4, 6,
     5, 1, 11, 7.
education: Children, 7th and 8th grade, 9th grade, 10th grade, High school graduate, 11th grade,
     12th grade no diploma, 5th or 6th grade, Less than 1st grade, Bachelors degree(BA AB BS),
     1st 2nd 3rd or 4th grade, Some college but no degree, Masters degree(MA MS MEng MEd MSW MBA),
     Associates degree-occup /vocational, Associates degree-academic program, Doctorate degree(PhD
     EdD), Prof school degree (MD DDS DVM LLB JD).
wage per hour: continuous.
enroll in edu inst last wk: Not in universe, High school, College or university.
marital stat: Never married, Married-civilian spouse present, Married-spouse absent, Separated,
     Divorced, Widowed, Married-A F spouse present.
major industry code: Not in universe or children, Entertainment, Social services, Agriculture,
     Education, Public administration, Manufacturing-durable goods, Manufacturing-nondurable goods,
     Wholesale trade, Retail trade, Finance insurance and real estate, Private household services,
     Business and repair services, Personal services except private HH, Construction, Medical
     except hospital, Other professional services, Transportation, Utilities and sanitary services,
     Mining, Communications, Hospital services, Forestry and fisheries, Armed Forces.
major occupation code: Not in universe, Professional specialty, Other service, Farming forestry and
     fishing, Sales, Adm support including clerical, Protective services, Handlers equip cleaners
     etc , Precision production craft & repair, Technicians and related support, Machine operators
     assmblrs & inspctrs, Transportation and material moving, Executive admin and managerial,
     Private household services, Armed Forces.
race: White, Black, Other, Amer Indian Aleut or Eskimo, Asian or Pacific Islander.
hispanic origin: Mexican (Mexicano), Mexican-American, Puerto Rican, Central or South American, All
     other, Other Spanish, Chicano, Cuban, Do not know, NA.
```

[1] UCI Machine Learning Repository Census-Income (KDD) Data Set: https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29

sex: Female, Male.

member of a labor union: Not in universe, No, Yes.

reason for unemployment: Not in universe, Re-entrant, Job loser - on layoff, New entrant, Job leaver, Other job loser.

full or part time employment stat: Children or Armed Forces, Full-time schedules, Unemployed part-time, Not in labor force, Unemployed full-time, PT for non-econ reasons usually FT, PT for econ reasons usually PT, PT for econ reasons usually FT.

capital gains: continuous.

capital losses: continuous.

dividends from stocks: continuous.

tax filer stat: Nonfiler, Joint one under 65 & one 65+, Joint both under 65, Single, Head of household, Joint both 65+.

region of previous residence: Not in universe, South, Northeast, West, Midwest, Abroad.

state of previous residence: Not in universe, Utah, Michigan, North Carolina, North Dakota, Virginia, Vermont, Wyoming, West Virginia, Pennsylvania, Abroad, Oregon, California, Iowa, Florida, Arkansas, Texas, South Carolina, Arizona, Indiana, Tennessee, Maine, Alaska, Ohio, Montana, Nebraska, Mississippi, District of Columbia, Minnesota, Illinois, Kentucky, Delaware, Colorado, Maryland, Wisconsin, New Hampshire, Nevada, New York, Georgia, Oklahoma, New Mexico, South Dakota, Missouri, Kansas, Connecticut, Louisiana, Alabama, Massachusetts, Idaho, New Jersey.

detailed household and family stat: Child <18 never marr not in subfamily, Other Rel <18 never marr child of subfamily RP, Other Rel <18 never marr not in subfamily, Grandchild <18 never marr child of subfamily RP, Grandchild <18 never marr not in subfamily, Secondary individual, In group quarters, Child under 18 of RP of unrel subfamily, RP of unrelated subfamily, Spouse of householder, Householder, Other Rel <18 never married RP of subfamily, Grandchild <18 never marr RP of subfamily, Child <18 never marr RP of subfamily, Child <18 ever marr not in subfamily, Other Rel <18 ever marr RP of subfamily, Child <18 ever marr RP of subfamily, Nonfamily householder, Child <18 spouse of subfamily RP, Other Rel <18 spouse of subfamily RP, Other Rel <18 ever marr not in subfamily, Grandchild <18 ever marr not in subfamily, Child 18+ never marr Not in a subfamily, Grandchild 18+ never marr not in subfamily, Child 18+ ever marr RP of subfamily, Other Rel 18+ never marr not in subfamily, Child 18+ never marr RP of subfamily, Other Rel 18+ ever marr RP of subfamily, Other Rel 18+ never marr RP of subfamily, Other Rel 18+ spouse of subfamily RP, Other Rel 18+ ever marr not in subfamily, Child 18+ ever marr Not in a subfamily, Grandchild 18+ ever marr not in subfamily, Child 18+ spouse of subfamily RP, Spouse of RP of unrelated subfamily, Grandchild 18+ ever marr RP of subfamily, Grandchild 18+ never marr RP of subfamily, Grandchild 18+ spouse of subfamily RP.

detailed household summary in household: Child under 18 never married, Other relative of householder, Nonrelative of householder, Spouse of householder, Householder, Child under 18 ever married, Group Quarters- Secondary individual, Child 18 or older.

instance weight: continuous.

migration code-change in msa: Not in universe, Nonmover, MSA to MSA, NonMSA to nonMSA, MSA to nonMSA, NonMSA to MSA, Abroad to MSA, Not identifiable, Abroad to nonMSA.

migration code-change in reg: Not in universe, Nonmover, Same county, Different county same state, Different state same division, Abroad, Different region, Different division same region.

migration code-move within reg: Not in universe, Nonmover, Same county, Different county same state, Different state in West, Abroad, Different state in Midwest, Different state in South, Different state in Northeast.

live in this house 1 year ago: Not in universe under 1 year old, Yes, No.

migration prev res in sunbelt: Not in universe, Yes, No.

num persons worked for employer: continuous.

family members under 18: Both parents present, Neither parent present, Mother only present, Father only present, Not in universe.

country of birth father: Mexico, United-States, Puerto-Rico, Dominican-Republic, Jamaica, Cuba, Portugal, Nicaragua, Peru, Ecuador, Guatemala, Philippines, Canada, Columbia, El-Salvador, Japan, England, Trinadad&Tobago, Honduras, Germany, Taiwan, Outlying-U S (Guam USVI etc), India, Vietnam, China, Hong Kong, Cambodia, France, Laos, Haiti, South Korea, Iran, Greece, Italy, Poland, Thailand, Yugoslavia, Holand-Netherlands, Ireland, Scotland, Hungary, Panama.

```
country of birth mother: India, Mexico, United-States, Puerto-Rico, Dominican-Republic, England,
     Honduras, Peru, Guatemala, Columbia, El-Salvador, Philippines, France, Ecuador, Nicaragua,
     Cuba, Outlying-U S (Guam USVI etc), Jamaica, South Korea, China, Germany, Yugoslavia, Canada,
     Vietnam, Japan, Cambodia, Ireland, Laos, Haiti, Portugal, Taiwan, Holand-Netherlands, Greece,
     Italy, Poland, Thailand, Trinadad&Tobago, Hungary, Panama, Hong Kong, Scotland, Iran.
country of birth self: United-States, Mexico, Puerto-Rico, Peru, Canada, South Korea, India, Japan,
     Haiti, El-Salvador, Dominican-Republic, Portugal, Columbia, England, Thailand, Cuba, Laos,
     Panama, China, Germany, Vietnam, Italy, Honduras, Outlying-U S (Guam USVI etc), Hungary,
     Philippines, Poland, Ecuador, Iran, Guatemala, Holand-Netherlands, Taiwan, Nicaragua, France,
     Jamaica, Scotland, Yugoslavia, Hong Kong, Trinadad&Tobago, Greece, Cambodia, Ireland.
citizenship: Native- Born in the United States, Foreign born- Not a citizen of U S , Native- Born
     in Puerto Rico or U S Outlying, Native- Born abroad of American Parent(s), Foreign born- U S
     citizen by naturalization.
own business or self employed: 0, 2, 1.
fill inc questionnaire for veteran's admin: Not in universe, Yes, No.
veterans benefits: 0, 2, 1.
weeks worked in year: continuous.
year: 94, 95.
total income: greater or lesser than 50000 (+50000,-50000)
```

A la vista de los atributos, observamos que hay datos cuantitativos continuos que debemos discretizar, así como datos numéricos que no son cuantitativos y que hay que hay transformar en factores.

Cargamos los datos en R.

```r
datos <- read.csv("census-income.csv", sep=",", strip.white=TRUE, stringsAsFactors=TRUE)
```

Imprimimos los valores de los atributos según como lo guarda R y lo comparamos con la lista anterior.

```r
for (aux in names(datos)){
  cat(aux, fill=TRUE)
  if(is.null(levels(datos[[aux]]))){
    cat(" numérico", fill=TRUE)
  } else{
  cat(levels(datos[[aux]]), fill=TRUE)
  }
}
```

```
## X
##  numérico
## age
##  numérico
## class.of.worker
##  Federal government  Local government  Never worked  Not in universe  Private
##  Self-employed-incorporated  Self-employed-not incorporated  State government
##  Without pay
## detailed.industry.recode
##  numérico
## detailed.ocupation.recode
##  numérico
## education
##  10th grade  11th grade  12th grade no diploma  1st 2nd 3rd or 4th grade
##  5th or 6th grade  7th and 8th grade  9th grade
##  Associates degree-academic program  Associates degree-occup /vocational
```

```
##  Bachelors degree(BA AB BS)  Children  Doctorate degree(PhD EdD)
##  High school graduate  Less than 1st grade
##  Masters degree(MA MS MEng MEd MSW MBA)  Prof school degree (MD DDS DVM LLB JD)
##  Some college but no degree
## wage.per.hour
##  numérico
## enroll.in.edu.inst.last.wk
##  College or university  High school  Not in universe
## marital.stat
##  Divorced  Married-A F spouse present  Married-civilian spouse present
##  Married-spouse absent  Never married  Separated  Widowed
## major.industry.code
##  Agriculture  Armed Forces  Business and repair services  Communications
##  Construction  Education  Entertainment  Finance insurance and real estate
##  Forestry and fisheries  Hospital services  Manufacturing-durable goods
##  Manufacturing-nondurable goods  Medical except hospital  Mining
##  Not in universe or children  Other professional services
##  Personal services except private HH  Private household services
##  Public administration  Retail trade  Social services  Transportation
##  Utilities and sanitary services  Wholesale trade
## major.occupation.code
##  Adm support including clerical  Armed Forces  Executive admin and managerial
##  Farming forestry and fishing  Handlers equip cleaners etc
##  Machine operators assmblrs & inspctrs  Not in universe  Other service
##  Precision production craft & repair  Private household services
##  Professional specialty  Protective services  Sales
##  Technicians and related support  Transportation and material moving
## race
##  Amer Indian Aleut or Eskimo  Asian or Pacific Islander  Black  Other  White
## hispanic.origin
##  All other  Central or South American  Chicano  Cuban  Do not know
##  Mexican-American  Mexican (Mexicano)  NA  Other Spanish  Puerto Rican
## sex
##  Female  Male
## member.of.a.labor.union
##  No  Not in universe  Yes
## reason.for.unemployment
##  Job leaver  Job loser - on layoff  New entrant  Not in universe
##  Other job loser  Re-entrant
## full.or.part.time.employment.stat
##  Children or Armed Forces  Full-time schedules  Not in labor force
##  PT for econ reasons usually FT  PT for econ reasons usually PT
##  PT for non-econ reasons usually FT  Unemployed full-time
##  Unemployed part- time
## capital.gains
##  numérico
## capital.losses
##  numérico
## dividends.from.stocks
##  numérico
## tax.filer.stat
##  Head of household  Joint both 65+  Joint both under 65
##  Joint one under 65 & one 65+  Nonfiler  Single
## region.of.previous.residence
```

```
##  Abroad  Midwest  Northeast  Not in universe  South  West
## state.of.previous.residence
##  ?  Abroad  Alabama  Alaska  Arizona  Arkansas  California  Colorado
##  Connecticut  Delaware  District of Columbia  Florida  Georgia  Idaho  Illinois
##  Indiana  Iowa  Kansas  Kentucky  Louisiana  Maine  Maryland  Massachusetts
##  Michigan  Minnesota  Mississippi  Missouri  Montana  Nebraska  Nevada
##  New Hampshire  New Jersey  New Mexico  New York  North Carolina  North Dakota
##  Not in universe  Ohio  Oklahoma  Oregon  Pennsylvania  South Carolina
##  South Dakota  Tennessee  Texas  Utah  Vermont  Virginia  West Virginia
##  Wisconsin  Wyoming
## detailed.household.and.family.stat
##  Child <18 ever marr not in subfamily  Child <18 ever marr RP of subfamily
##  Child <18 never marr not in subfamily  Child <18 never marr RP of subfamily
##  Child <18 spouse of subfamily RP  Child 18+ ever marr Not in a subfamily
##  Child 18+ ever marr RP of subfamily  Child 18+ never marr Not in a subfamily
##  Child 18+ never marr RP of subfamily  Child 18+ spouse of subfamily RP
##  Child under 18 of RP of unrel subfamily
##  Grandchild <18 ever marr not in subfamily
##  Grandchild <18 never marr child of subfamily RP
##  Grandchild <18 never marr not in subfamily
##  Grandchild <18 never marr RP of subfamily
##  Grandchild 18+ ever marr not in subfamily
##  Grandchild 18+ ever marr RP of subfamily
##  Grandchild 18+ never marr not in subfamily
##  Grandchild 18+ never marr RP of subfamily
##  Grandchild 18+ spouse of subfamily RP  Householder  In group quarters
##  Nonfamily householder  Other Rel <18 ever marr not in subfamily
##  Other Rel <18 ever marr RP of subfamily
##  Other Rel <18 never marr child of subfamily RP
##  Other Rel <18 never marr not in subfamily
##  Other Rel <18 never married RP of subfamily
##  Other Rel <18 spouse of subfamily RP  Other Rel 18+ ever marr not in subfamily
##  Other Rel 18+ ever marr RP of subfamily
##  Other Rel 18+ never marr not in subfamily
##  Other Rel 18+ never marr RP of subfamily  Other Rel 18+ spouse of subfamily RP
##  RP of unrelated subfamily  Secondary individual  Spouse of householder
##  Spouse of RP of unrelated subfamily
## detailed.household.summary.in.household
##  Child 18 or older  Child under 18 ever married  Child under 18 never married
##  Group Quarters- Secondary individual  Householder  Nonrelative of householder
##  Other relative of householder  Spouse of householder
## instance.weight
##  numérico
## migration.code.change.in.msa
##  ?  Abroad to MSA  Abroad to nonMSA  MSA to MSA  MSA to nonMSA  Nonmover
##  NonMSA to MSA  NonMSA to nonMSA  Not identifiable  Not in universe
## migration.code.change.in.reg
##  ?  Abroad  Different county same state  Different division same region
##  Different region  Different state same division  Nonmover  Not in universe
##  Same county
## migration.code.move.within.reg
##  ?  Abroad  Different county same state  Different state in Midwest
##  Different state in Northeast  Different state in South
##  Different state in West  Nonmover  Not in universe  Same county
```

```
## live.in.this.house.1.year.ago
##  No  Not in universe under 1 year old  Yes
## migration.prev.res.in.sunbelt
##  ?  No  Not in universe  Yes
## num.persons.worked.for.employer
##  numérico
## family.members.under.18
##  Both parents present  Father only present  Mother only present
##  Neither parent present  Not in universe
## country.of.birth.father
##  ?  Cambodia  Canada  China  Columbia  Cuba  Dominican-Republic  Ecuador
##  El-Salvador  England  France  Germany  Greece  Guatemala  Haiti
##  Holand-Netherlands  Honduras  Hong Kong  Hungary  India  Iran  Ireland  Italy
##  Jamaica  Japan  Laos  Mexico  Nicaragua  Outlying-U S (Guam USVI etc)  Panama
##  Peru  Philippines  Poland  Portugal  Puerto-Rico  Scotland  South Korea
##  Taiwan  Thailand  Trinadad&Tobago  United-States  Vietnam  Yugoslavia
## country.of.birth.mother
##  ?  Cambodia  Canada  China  Columbia  Cuba  Dominican-Republic  Ecuador
##  El-Salvador  England  France  Germany  Greece  Guatemala  Haiti
##  Holand-Netherlands  Honduras  Hong Kong  Hungary  India  Iran  Ireland  Italy
##  Jamaica  Japan  Laos  Mexico  Nicaragua  Outlying-U S (Guam USVI etc)  Panama
##  Peru  Philippines  Poland  Portugal  Puerto-Rico  Scotland  South Korea
##  Taiwan  Thailand  Trinadad&Tobago  United-States  Vietnam  Yugoslavia
## country.of.birth.self
##  ?  Cambodia  Canada  China  Columbia  Cuba  Dominican-Republic  Ecuador
##  El-Salvador  England  France  Germany  Greece  Guatemala  Haiti
##  Holand-Netherlands  Honduras  Hong Kong  Hungary  India  Iran  Ireland  Italy
##  Jamaica  Japan  Laos  Mexico  Nicaragua  Outlying-U S (Guam USVI etc)  Panama
##  Peru  Philippines  Poland  Portugal  Puerto-Rico  Scotland  South Korea
##  Taiwan  Thailand  Trinadad&Tobago  United-States  Vietnam  Yugoslavia
## citizenship
##  Foreign born- Not a citizen of U S
##  Foreign born- U S citizen by naturalization
##  Native- Born abroad of American Parent(s)
##  Native- Born in Puerto Rico or U S Outlying  Native- Born in the United States
## own.business.or.self.employed
##  numérico
## fill.inc.questionnaire.for.veteran.s.admin
##  No  Not in universe  Yes
## veterans.benefits
##  numérico
## weeks.worked.in.year
##  numérico
## year
##  numérico
## total.income
##  - 50000.  50000+.
```

Tenemos un atirbuto `instance.weight` que indica el número de individuos del censo que comparten los mismos valores de los atributos, es decir, el peso de la instancia en la muestra. Esto significa que debemos hacer reglas de asociación con pesos para poder reflejar la realidad de los datos, ya que así damos importancia proporcional a cada instancia según el número de individuos que representa.

Extraemos los pesos de los datos y el primer atributo que indica simplemente el número de la instancia porque no deben aparecer en las reglas de asociación.

```
pesos <- subset(datos, select=instance.weight)
datos <- subset(datos, select=-c(X, instance.weight))
```

Observamos también que tenemos datos numéricos que en realidad son cualitativos, así que los transformamos en factores.

```
afactor <- c("detailed.industry.recode", "detailed.ocupation.recode",
             "own.business.or.self.employed", "veterans.benefits", "year")
datos[afactor] <- as.data.frame(lapply(datos[afactor], as.factor))
```

El resto de los datos numéricos se corresponden a atributos cuantitativos continuos como la edad. Para poder hacer minería de reglas de asociación necesitamos transformarlos en cualitativos discretizándolos. Discretizamos la edad con tantos *bines* como la raíz cuadrada del número de datos. Para el resto, observamos que el hecho de ser 0 tiene un significado especial: nada de sueldo por hora, ninguna ganancia o pérdida de capital, ningún dividendo, ningún trabajador o ninguna semana por año. Por lo tanto, queremos que el 0 sea una categoría independiente en estos atributos. Discretizamos mediante un método de *clustering* con $k$-medias de manera que no tomamos los extremos inferiores de los intervalos. De esta manera, los valores 0 se discretizan como NA.

```
datos <- discretizeDF(datos, methods=list(
  age=list(method="interval", breaks=sqrt(length(datos$age))),
  wage.per.hour=list(method ="cluster", include.lowest=FALSE, right=TRUE),
  capital.gains=list(method="cluster", include.lowest=FALSE, right=TRUE),
  capital.losses=list(method="cluster", include.lowest=FALSE, right=TRUE),
  dividends.from.stocks=list(method="cluster", include.lowest=FALSE, right=TRUE),
  num.persons.worked.for.employer=list(method="cluster", include.lowest=FALSE, right=TRUE),
  weeks.worked.in.year=list(method="cluster", include.lowest=FALSE, right=TRUE)))
```

Queremos convertir los NA en 0. Para ello, cambiamos temporalmente los datos a caracteres, transformamos NA por 0 y deshacemos el cambio.

```
datos <- as.data.frame(lapply(datos, as.character))
datos[is.na(datos)] <- "0"
datos <- as.data.frame(lapply(datos, as.factor))
```

Observamos que hay nueve atributos que tienen valores faltantes para algunas instancias. Si eliminásemos todas cuyo valor para alguno de estos atributos falta, perderíamos alrededor del 47% de los datos. Por el contrario, si solo eliminásemos las instancias para los atributos que estemos considerando en cada momento en las reglas, estas describirían cada vez un conjunto distinto de datos, con lo que no se podrían comparar entre sí. Entre perder sobre el 47% de los datos o nueve atributos, priorizaremos lo segundo. En cualquier caso, definimos la siguiente función a la que se le pasan los atributos sobre los que se quiere construir reglas de asociación y elimina las instancias con valores faltantes. Tenemos en cuenta que debemos eliminar también el peso correspondiente a cada instancia eliminada.

```
eliminar_faltantes <- function(atributos){
  datos <- cbind(datos, pesos)
  bool <- atributos %in% c("hispanic.origin",
                           "state.of.previous.residence",
                           "detailed.household.and.family.stat",
                           "migration.code.change.in.msa",
                           "migration.code.change.in.reg",
```

```
                           "migration.prev.res.in.sunbelt",
                           "country.of.birth.father",
                           "country.of.birth.mother",
                           "country.of.birth.self")
  for (aux in atributos[bool]){
    datos <- droplevels(subset(datos, datos[aux]!=" ?"))
    datos <- droplevels(subset(datos, datos[[aux]]!=" NA"))
  }
  pesos_lista <<- datos$instance.weight
  datos <<- subset.data.frame(datos, select=-instance.weight)
}
```

Con esto quedan preprocesados los datos para cualquier tipo de regla de asociación.

## Reglas de asociación

Para minar las reglas de asociación utilizaremos el algoritmo Eclat con pesos. Para ello, definimos la siguiente función que permite calcularlas para los atributos que queramos fijando el soporte y la confianza mínimos así como la longitud mínima y máxima.

```
reglas_weclat <- function(atributos, sop, conf, minl=1, maxl=5){
  # Eliminamos las instancias con valores de atributos faltantes si es el caso.
  eliminar_faltantes(atributos)

  # Preparamos la entrada para el algoritmo Eclat con pesos.
  transacciones <- as(datos[atributos], "transactions")

  # Añadimos los pesos.
  transacciones@itemsetInfo$weight <- NULL
  transacciones@itemsetInfo$weight <- pesos_lista

  # Parámetros del algoritmo Eclat con pesos.
  aparametros <- list(support=sop, minlen=minl, maxlen=maxl)

  # Calculamos los itemsets frecuentes con el algoritmo Eclat.
  itemsets <- weclat(transacciones, aparametros, control=list(verbose=FALSE))
  itemsets.df <- data.frame(itemsets=labels(itemsets), itemsets@quality)

  # Inducción de las reglas de asociación.
  reglas <- ruleInduction(itemsets, transacciones, confidence=conf)

  # Devolvemos las reglas calculadas que no sean redundantes.
  return(reglas[!is.redundant(reglas)])
}
```

Para poder aplicar la función anterior y visualizar las reglas necesitaremos los paquetes `arules` y `arulesViz` que importamos de la siguiente manera.

```
library(arules)
library(arulesViz)
```

Aun sin tener en cuenta los atributos con datos faltantes, tenemos 32 atributos. Si utilizamos varios de ellos, a poco que disminuyamos el soporte mínimo obtendremos demasiadas reglas de asociación independientemente de que la confianza mínima escogida sea alta. Además, casi todos los atributos tienen un valor mayoritario respecto al resto lo que imposibilita obtener reglas con alto soporte y valores de atributos variados por el simple hecho de que esos valores no tienen ese soporte en la muestra. Podemos observar este fenómeno mediante un resumen frecuentista de los datos.

```
summary(datos)
```

```
##           age                                class.of.worker
##  [33.8,34)  :  3489    Not in universe               :100245
##  [34.8,35)  :  3450    Private                       : 72028
##  [35.8,36)  :  3353    Self-employed-not incorporated:  8445
##  [30.8,31)  :  3351    Local government              :  7784
##  [32.8,33)  :  3340    State government              :  4227
##  [4.83,5.03):  3332    Self-employed-incorporated    :  3265
##  (Other)    :179208   (Other)                        :  3529
##  detailed.industry.recode detailed.ocupation.recode
##  0      :100684           0      :100684
##  33     : 17070           2      :  8756
##  43     :  8283           26     :  7887
##  4      :  5984           19     :  5413
##  42     :  4683           29     :  5105
##  45     :  4482           36     :  4145
##  (Other): 58337           (Other): 67533
##                          education              wage.per.hour
##   High school graduate      :48407   (0,382]         :   181
##   Children                  :47422   (1.36e+03,1e+04]:  2117
##   Some college but no degree:27820   (382,1.36e+03]  :  9006
##   Bachelors degree(BA AB BS):19865   0               :188219
##   7th and 8th grade         : 8007
##   10th grade                : 7557
##  (Other)                    :40445
##           enroll.in.edu.inst.last.wk                          marital.stat
##   College or university:  5688       Divorced                       :12710
##   High school          :  6892       Married-A F spouse present     :  665
##   Not in universe      :186943       Married-civilian spouse present:84222
##                                      Married-spouse absent          : 1518
##                                      Never married                  :86485
##                                      Separated                      : 3460
##                                      Widowed                        :10463
##                            major.industry.code
##   Not in universe or children      :100684
##   Retail trade                     : 17070
##   Manufacturing-durable goods      :  9015
##   Education                        :  8283
##   Manufacturing-nondurable goods   :  6897
##   Finance insurance and real estate:  6145
##  (Other)                           : 51429
##                       major.occupation.code
##   Not in universe              :100684
##   Adm support including clerical: 14837
##   Professional specialty        : 13940
##   Executive admin and managerial: 12495
```

```
##   Other service                : 12099
##   Sales                        : 11783
##  (Other)                       : 33685
##                           race                         hispanic.origin
##   Amer Indian Aleut or Eskimo:  2251    All other               :171907
##   Asian or Pacific Islander  :  5835    Mexican-American        :  8079
##   Black                      : 20415    Mexican (Mexicano)      :  7234
##   Other                      :  3657    Central or South American:  3895
##   White                      :167365    Puerto Rican            :  3313
##                                         Other Spanish           :  2485
##                                        (Other)                  :  2610
##       sex            member.of.a.labor.union          reason.for.unemployment
##   Female:103984    No             : 16034      Job leaver           :   598
##   Male  : 95539    Not in universe:180459      Job loser - on layoff:   976
##                    Yes            :  3030      New entrant          :   439
##                                                Not in universe      :193453
##                                                Other job loser      :  2038
##                                                Re-entrant           :  2019
##
##                    full.or.part.time.employment.stat
##   Children or Armed Forces          :123769
##   Full-time schedules               : 40736
##   Not in labor force                : 26808
##   PT for non-econ reasons usually FT:  3322
##   Unemployed full-time              :  2311
##   PT for econ reasons usually PT    :  1209
##  (Other)                            :  1368
##              capital.gains                 capital.losses
##  (0,5.75e+03]       :  3841   (0,900]            :   101
##  (5.57e+04,1e+05]   :   390   (2.18e+03,4.61e+03]:   710
##  (5.75e+03,5.57e+04]:  3148   (900,2.18e+03]     :  3095
##  0                  :192144   0                  :195617
##
##
##
##          dividends.from.stocks                          tax.filer.stat
##  (0,6.43e+03]       : 19753     Head of household           : 7426
##  (3.73e+04,1e+05]   :   113     Joint both 65+              : 8332
##  (6.43e+03,3.73e+04]:  1275     Joint both under 65         :67383
##  0                  :178382     Joint one under 65 & one 65+: 3867
##                                 Nonfiler                    :75094
##                                 Single                      :37421
##
##    region.of.previous.residence   state.of.previous.residence
##   Abroad         :   530         Not in universe:183750
##   Midwest        :  3575         California     :  1714
##   Northeast      :  2705         Utah           :  1063
##   Not in universe:183750         Florida        :   849
##   South          :  4889         North Carolina :   812
##   West           :  4074         ?              :   708
##                                 (Other)         : 10627
##                         detailed.household.and.family.stat
##   Householder                             :53248
##   Child <18 never marr not in subfamily  :50326
```

```
##   Spouse of householder                 :41695
##   Nonfamily householder                 :22213
##   Child 18+ never marr Not in a subfamily:12030
##   Secondary individual                  : 6122
##  (Other)                                :13889
##            detailed.household.summary.in.household
##   Householder                 :75475
##   Child under 18 never married :50426
##   Spouse of householder       :41709
##   Child 18 or older           :14430
##   Other relative of householder: 9703
##   Nonrelative of householder   : 7601
##  (Other)                       :  179
##    migration.code.change.in.msa               migration.code.change.in.reg
##   ?               :99696         ?                           :99696
##   Nonmover        :82538         Nonmover                    :82538
##   MSA to MSA      :10601         Same county                 : 9812
##   NonMSA to nonMSA: 2811         Different county same state: 2797
##   Not in universe : 1516         Not in universe             : 1516
##   MSA to nonMSA   :  790         Different region            : 1178
##  (Other)          : 1571        (Other)                      : 1986
##               migration.code.move.within.reg
##   ?                          :99696
##   Nonmover                   :82538
##   Same county                : 9812
##   Different county same state: 2797
##   Not in universe            : 1516
##   Different state in South   :  973
##  (Other)                     : 2191
##                   live.in.this.house.1.year.ago  migration.prev.res.in.sunbelt
##   No                              : 15773        ?              :99696
##   Not in universe under 1 year old:101212        No             : 9987
##   Yes                             : 82538        Not in universe:84054
##                                                  Yes            : 5786
##
##
##
##  num.persons.worked.for.employer            family.members.under.18
##  (0,1.38]   :23109              Both parents present  : 38983
##  (1.38,3.98]:23506              Father only present   :  1883
##  (3.98,6]   :56925              Mother only present   : 12772
##  0          :95983              Neither parent present:  1653
##                                 Not in universe       :144232
##
##
##    country.of.birth.father   country.of.birth.mother    country.of.birth.self
##   United-States:159163      United-States:160479      United-States:176989
##   Mexico       : 10008      Mexico       :  9781      Mexico       :  5767
##   ?            :  6713      ?            :  6119      ?            :  3393
##   Puerto-Rico  :  2680      Puerto-Rico  :  2473      Puerto-Rico  :  1400
##   Italy        :  2212      Italy        :  1844      Germany      :   851
##   Canada       :  1380      Canada       :  1451      Philippines  :   845
##  (Other)       : 17367     (Other)       : 17376     (Other)       : 10278
##                                       citizenship
```

```
##   Foreign born- Not a citizen of U S         : 13401
##   Foreign born- U S citizen by naturalization:  5855
##   Native- Born abroad of American Parent(s)  :  1756
##   Native- Born in Puerto Rico or U S Outlying:  1519
##   Native- Born in the United States          :176992
##
##
##  own.business.or.self.employed fill.inc.questionnaire.for.veteran.s.admin
##  0:180672                       No             :  1593
##  1:  2698                       Not in universe:197539
##  2: 16153                       Yes            :   391
##
##
##
##
##  veterans.benefits  weeks.worked.in.year year          total.income
##  0: 47409         (0,15.1]   : 8226     94:99827    - 50000.:187141
##  1:  1984         (15.1,40.5]:16027     95:99696    50000+. : 12382
##  2:150130         (40.5,52]  :79287
##                   0          :95983
##
##
##
```

Por poner solo un ejemplo, el atributo `citizenship` tiene un valor que abarca una inmensa mayoría de las instancias, a saber, *Native- Born in the United States*. Podemos calcular las frecuencias relativas de los valores.

```
table(datos$citizenship)/sum(table(datos$citizenship))
```

```
##
##          Foreign born- Not a citizen of U S
##                                 0.067165189
##  Foreign born- U S citizen by naturalization
##                                 0.029344988
##    Native- Born abroad of American Parent(s)
##                                 0.008800990
##  Native- Born in Puerto Rico or U S Outlying
##                                 0.007613157
##            Native- Born in the United States
##                                 0.887075675
```

Aunque depende en parte del peso de las instancias, vemos que si fijásemos un soporte mayor que 0.06712, solo podríamemos obtener reglas para ciudadanos nativos estadounideneses. Esto nos da una idea de lo bajo que debemos fijar el soporte mínimo para obtener reglas que involucren diferentes valores de las variables.

La consideraciones anteriores no contemplan el hecho de que la mayoría de las reglas de asociación tienen como consecuente el valor mayoritario de un atributo, con lo que es difícil visualizar asociaciones que no terminen en los mismos valores. Obtendremos reglas con consecuentes variados simplemente tomando subconjuntos de las reglas que los contengan.

Una vez escogidos el soporte y la confianza mínimos, generaremos las reglas y las filtraremos y representaremos algunos subconjuntos pequeños mediante coordenadas paralelas. Estos diagramas están formados por flechas que conectan los antecedentes y terminan en el consecuente. El grosor de la flecha indica el soporte y la saturación de color la confianza. Utilizaremos la opción `reorder=TRUE` que aplica un método heurístico para minimizar el número de intersecciones de las flechas.

```
paracoord_reglas <- function(reglas){
  plot(reglas, method="paracoord", reorder=TRUE)
}
```

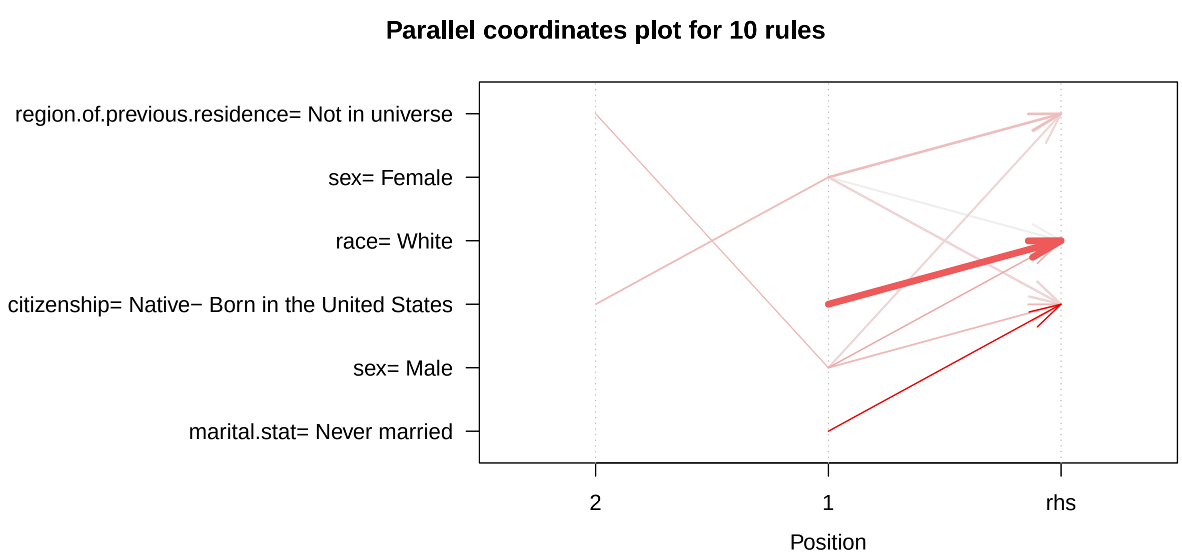
También mostraremos las reglas de manera explícita junto a su soporte y confianza, para lo que definimos la siguiente función.

```
print_reglas <- function(reglas){
  reglas.df <- data.frame(rules=labels(reglas), reglas@quality)
  reglas.df$rules <- as.character(reglas.df$rules)
  reglas.df <- reglas.df[order(reglas.df$support),]
  for (i in 1:length(reglas.df$rules)){
    cat(paste(reglas.df$rules[i],
              round(reglas.df$support[i], 4),
              round(reglas.df$confidence[i], 4)), fill=TRUE)
  }
}
```

### Reglas de asociación de tipo sociodemográfico

Calcularemos un primer conjunto de reglas de asociación relativas a atributos de caracter sociodemográfico como puede ser la edad, el género, la etnia o la ciudadanía sin tener en cuenta atributos de tipo puramente laboral o económico.

```
reglas <- reglas_weclat(c("age", "race", "sex", "marital.stat",
                             "region.of.previous.residence",
                             "detailed.household.summary.in.household",
                             "citizenship"), sop=0.4, conf=0.8)
paracoord_reglas(reglas[1:10])
```



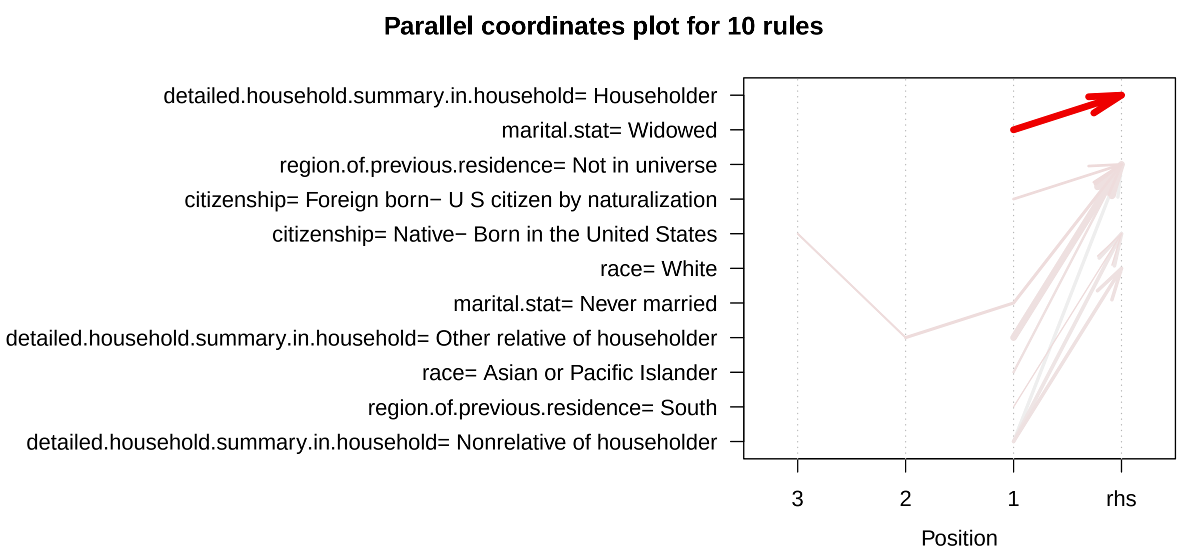
```r
print_reglas(reglas[1:10])
```

```
## {sex= Male,region.of.previous.residence= Not in universe} => {citizenship= Native-
Born in the United States} 0.3924 0.8913
## {marital.stat= Never married} => {citizenship= Native- Born in the United States}
0.3976 0.9173
## {sex= Male} => {race= White} 0.4052 0.8461
## {sex= Male} => {citizenship= Native- Born in the United States} 0.4254 0.8885
## {sex= Female,citizenship= Native- Born in the United States} =>
{region.of.previous.residence= Not in universe} 0.4268 0.9246
## {sex= Female} => {race= White} 0.4337 0.8321
## {sex= Male} => {region.of.previous.residence= Not in universe} 0.4402 0.9193
## {sex= Female} => {citizenship= Native- Born in the United States} 0.4616 0.8858
## {sex= Female} => {region.of.previous.residence= Not in universe} 0.4807 0.9224
## {citizenship= Native- Born in the United States} => {race= White} 0.7604 0.8572
```

Obtenemos resultados esperados, reglas con alto soporte y confianza que indican que la mayoría de los datos se refieren a ciudadanos nativos estadounidenses blancos que no han cambiado su región de residencia. Las reglas que involucran el género tienen menor soporte porque la proporción aproximada en los datos es la mitad para masculino y femenino.

Probamos ahora con un soporte muy pequeño del 0.02 y observamos que, efectivamente, obtenemos reglas con valores más variados para los atributos.

```r
reglas<- reglas_weclat(c("age", "race", "sex", "marital.stat",
                             "region.of.previous.residence",
                             "detailed.household.summary.in.household",
                             "citizenship"), sop=0.02, conf=0.8)
paracoord_reglas(reglas[1:10])
```

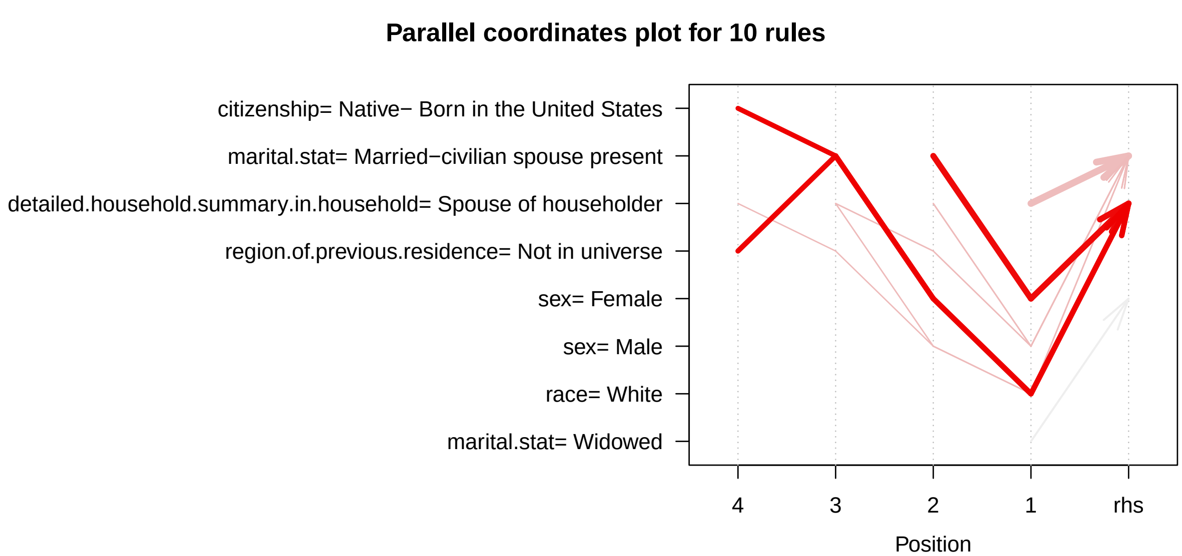



```r
print_reglas(reglas[1:10])
```

```
## {region.of.previous.residence= South} => {citizenship= Native- Born in the United
States} 0.0221 0.9037
## {marital.stat= Never married,detailed.household.summary.in.household= Other relative
of householder,citizenship= Native- Born in the United States} =>
{region.of.previous.residence= Not in universe} 0.0253 0.9356
## {race= Asian or Pacific Islander} => {region.of.previous.residence= Not in universe}
0.0266 0.9102
## {citizenship= Foreign born- U S citizen by naturalization} =>
{region.of.previous.residence= Not in universe} 0.0278 0.9484
## {marital.stat= Never married,detailed.household.summary.in.household= Other relative
of householder} => {region.of.previous.residence= Not in universe} 0.0296 0.9252
## {detailed.household.summary.in.household= Nonrelative of householder} =>
{region.of.previous.residence= Not in universe} 0.0306 0.802
## {detailed.household.summary.in.household= Nonrelative of householder} => {race= White}
0.0313 0.8216
## {detailed.household.summary.in.household= Nonrelative of householder} => {citizenship=
Native- Born in the United States} 0.032 0.8407
## {detailed.household.summary.in.household= Other relative of householder} =>
{region.of.previous.residence= Not in universe} 0.0442 0.9095
## {marital.stat= Widowed} => {detailed.household.summary.in.household= Householder}
0.0451 0.8591
```

Extraemos del subconjunto de reglas calculadas aquellas que no terminan en valores mayoritarios de los atributos.

```r
reglas.sub <- subset(reglas, subset= !(
  rhs %oin% c("marital.stat= Never married",
              "citizenship= Native- Born in the United States", "race= White",
              "detailed.household.summary.in.household= Householder",
              "region.of.previous.residence= Not in universe")))
paracoord_reglas(reglas.sub[1:10])
```



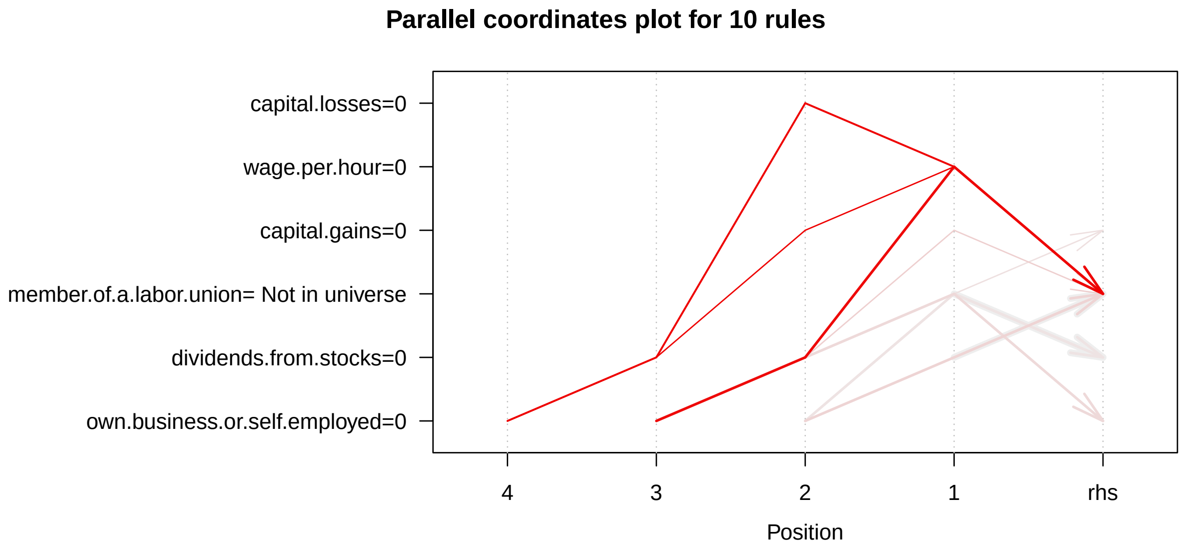
```r
print_reglas(reglas.sub[1:10])
```

```
## {race= White,sex= Male,region.of.previous.residence= Not in
universe,detailed.household.summary.in.household= Spouse of householder} =>
{marital.stat= Married-civilian spouse present} 0.0229 0.998
## {race= White,sex= Male,detailed.household.summary.in.household= Spouse of householder}
=> {marital.stat= Married-civilian spouse present} 0.0246 0.9976
## {sex= Male,region.of.previous.residence= Not in
universe,detailed.household.summary.in.household= Spouse of householder} =>
{marital.stat= Married-civilian spouse present} 0.0267 0.9979
## {sex= Male,detailed.household.summary.in.household= Spouse of householder} =>
{marital.stat= Married-civilian spouse present} 0.0288 0.9974
## {marital.stat= Widowed} => {sex= Female} 0.0437 0.833
## {race= White,sex= Female,marital.stat= Married-civilian spouse present,citizenship=
Native- Born in the United States} => {detailed.household.summary.in.household= Spouse of
householder} 0.143 0.8593
## {race= White,sex= Female,marital.stat= Married-civilian spouse
present,region.of.previous.residence= Not in universe} =>
{detailed.household.summary.in.household= Spouse of householder} 0.1504 0.8551
## {race= White,sex= Female,marital.stat= Married-civilian spouse present} =>
{detailed.household.summary.in.household= Spouse of householder} 0.1595 0.853
## {sex= Female,marital.stat= Married-civilian spouse present} =>
{detailed.household.summary.in.household= Spouse of householder} 0.1773 0.8436
## {detailed.household.summary.in.household= Spouse of householder} => {marital.stat=
Married-civilian spouse present} 0.2061 0.9859
```

Visualizando parte de las reglas, observamos que ahora tenemos información más interesante.

### Reglas de asociación de tipo laboral-económico

```
reglas <- reglas_weclat(c("class.of.worker", "education", "wage.per.hour",
                          "enroll.in.edu.inst.last.wk",
                          "member.of.a.labor.union",
                          "reason.for.unemployment",
                          "full.or.part.time.employment.stat",
                          "capital.gains", "capital.losses",
                          "dividends.from.stocks", "tax.filer.stat",
                          "num.persons.worked.for.employer",
                          "own.business.or.self.employed"), sop=0.7, conf=0.8)
paracoord_reglas(reglas[1:10])
```




```
print_reglas(reglas[1:10])
```

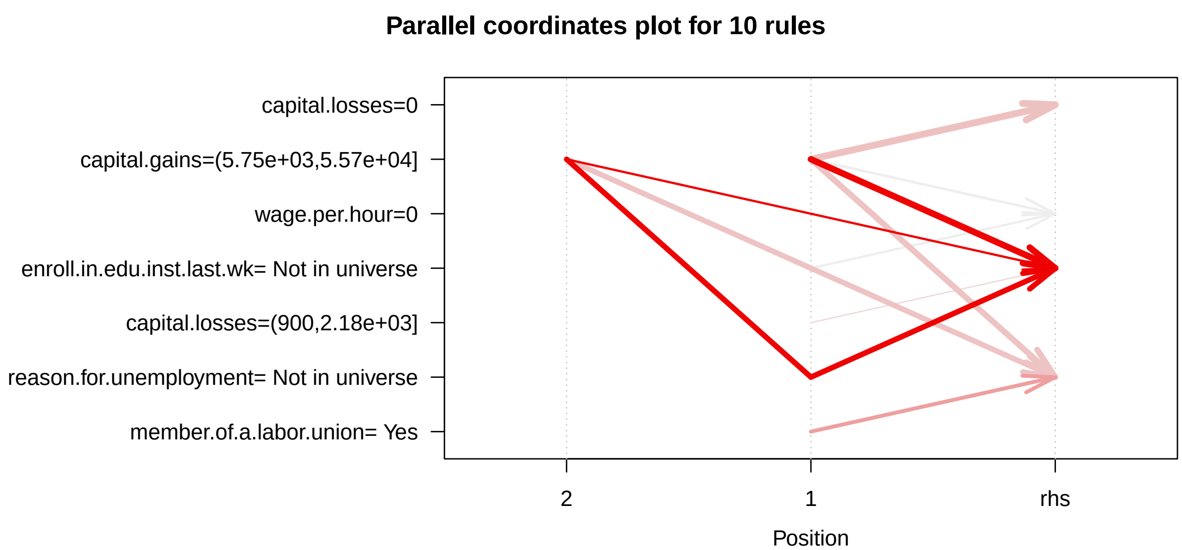
```
##
{wage.per.hour=0,capital.gains=0,dividends.from.stocks=0,own.business.or.self.employed=0}
=> {member.of.a.labor.union= Not in universe} 0.7276 0.969
## {capital.gains=0,dividends.from.stocks=0,own.business.or.self.employed=0} =>
{member.of.a.labor.union= Not in universe} 0.7276 0.921
## {member.of.a.labor.union= Not in
universe,dividends.from.stocks=0,own.business.or.self.employed=0} => {capital.gains=0}
0.7276 0.9741
##
{wage.per.hour=0,capital.losses=0,dividends.from.stocks=0,own.business.or.self.employed=0}
=> {member.of.a.labor.union= Not in universe} 0.7365 0.9684
## {dividends.from.stocks=0,own.business.or.self.employed=0} => {member.of.a.labor.union=
Not in universe} 0.747 0.9192
## {member.of.a.labor.union= Not in universe,own.business.or.self.employed=0} =>
{dividends.from.stocks=0} 0.747 0.903
## {member.of.a.labor.union= Not in universe,dividends.from.stocks=0} =>
{own.business.or.self.employed=0} 0.747 0.9183
```

```
## {wage.per.hour=0,dividends.from.stocks=0,own.business.or.self.employed=0} =>
{member.of.a.labor.union= Not in universe} 0.747 0.9675
## {dividends.from.stocks=0} => {member.of.a.labor.union= Not in universe} 0.8135 0.9099
## {member.of.a.labor.union= Not in universe} => {dividends.from.stocks=0} 0.8135 0.8994
```

Con un valor alto para el soporte, tenemos las reglas de asociación esperadas, que se refieren mayoritariamente a ciudadanos que no trabajan y por lo tanto, no tienen sueldo, ni ganancias ni pertencen a un sindicato.

Buscamos obtener reglas que asocien valores menos frecuentes por lo que consideramos un soporte mucho menor.

```
reglas <- reglas_weclat(c("class.of.worker", "wage.per.hour",
                          "enroll.in.edu.inst.last.wk", "member.of.a.labor.union",
                          "reason.for.unemployment",
                          "full.or.part.time.employment.stat", "capital.gains",
                          "capital.losses", "dividends.from.stocks",
                          "num.persons.worked.for.employer",
                          "own.business.or.self.employed", "weeks.worked.in.year"),
                        sop=0.015, conf=0.75)
paracoord_reglas(reglas[1:10])
```



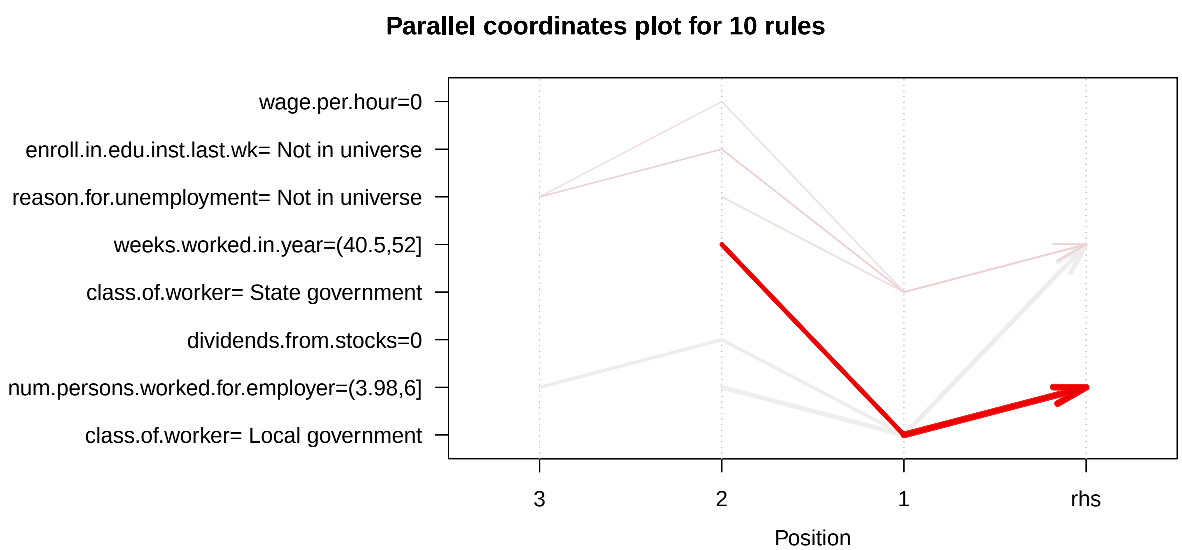

```
print_reglas(reglas[1:10])
```

```
## {capital.losses=(900,2.18e+03]} => {enroll.in.edu.inst.last.wk= Not in universe}
0.0147 0.9473
## {enroll.in.edu.inst.last.wk= Not in universe,capital.gains=(5.75e+03,5.57e+04]} =>
{wage.per.hour=0} 0.0149 0.9479
## {wage.per.hour=0,capital.gains=(5.75e+03,5.57e+04]} => {enroll.in.edu.inst.last.wk=
Not in universe} 0.0149 0.9946
## {capital.gains=(5.75e+03,5.57e+04]} => {wage.per.hour=0} 0.015 0.9476
## {member.of.a.labor.union= Yes} => {reason.for.unemployment= Not in universe} 0.0152 1
## {reason.for.unemployment= Not in universe,capital.gains=(5.75e+03,5.57e+04]} =>
{enroll.in.edu.inst.last.wk= Not in universe} 0.0155 0.9945
```

```
## {enroll.in.edu.inst.last.wk= Not in universe,capital.gains=(5.75e+03,5.57e+04]} =>
{reason.for.unemployment= Not in universe} 0.0155 0.9882
## {capital.gains=(5.75e+03,5.57e+04]} => {reason.for.unemployment= Not in universe}
0.0156 0.9879
## {capital.gains=(5.75e+03,5.57e+04]} => {enroll.in.edu.inst.last.wk= Not in universe}
0.0157 0.9943
## {capital.gains=(5.75e+03,5.57e+04]} => {capital.losses=0} 0.0158 1
```

Conseguimos obtener reglas que involucran las ganancias, las pérdidas o la pertenencia a un sindicato.

```r
reglas.sub <- subset(reglas, subset= !(rhs %oin% c(
  "class.of.worker= Not in universe", "wage.per.hour=0",
  "enroll.in.edu.inst.last.wk= Not in universe",
  "member.of.a.labor.union= Not in universe",
  "reason.for.unemployment= Not in universe",
  "full.or.part.time.employment.stat= Children or Armed Forces",
  "capital.gains=0", "capital.losses=0", "dividends.from.stocks=0",
  "num.persons.worked.for.employer=0", "own.business.or.self.employed=0",
  "weeks.worked.in.year=0")))
paracoord_reglas(reglas.sub[1:10])
```




```r
print_reglas(reglas.sub[1:10])
```

```
## {class.of.worker= State government,wage.per.hour=0,reason.for.unemployment= Not in
universe} => {weeks.worked.in.year=(40.5,52]} 0.0156 0.8384
## {class.of.worker= State government,wage.per.hour=0} =>
{weeks.worked.in.year=(40.5,52]} 0.0156 0.8183
## {class.of.worker= State government,enroll.in.edu.inst.last.wk= Not in
universe,reason.for.unemployment= Not in universe} => {weeks.worked.in.year=(40.5,52]}
0.0164 0.8625
## {class.of.worker= State government,enroll.in.edu.inst.last.wk= Not in universe} =>
{weeks.worked.in.year=(40.5,52]} 0.0165 0.8442
## {class.of.worker= State government,reason.for.unemployment= Not in universe} =>
```

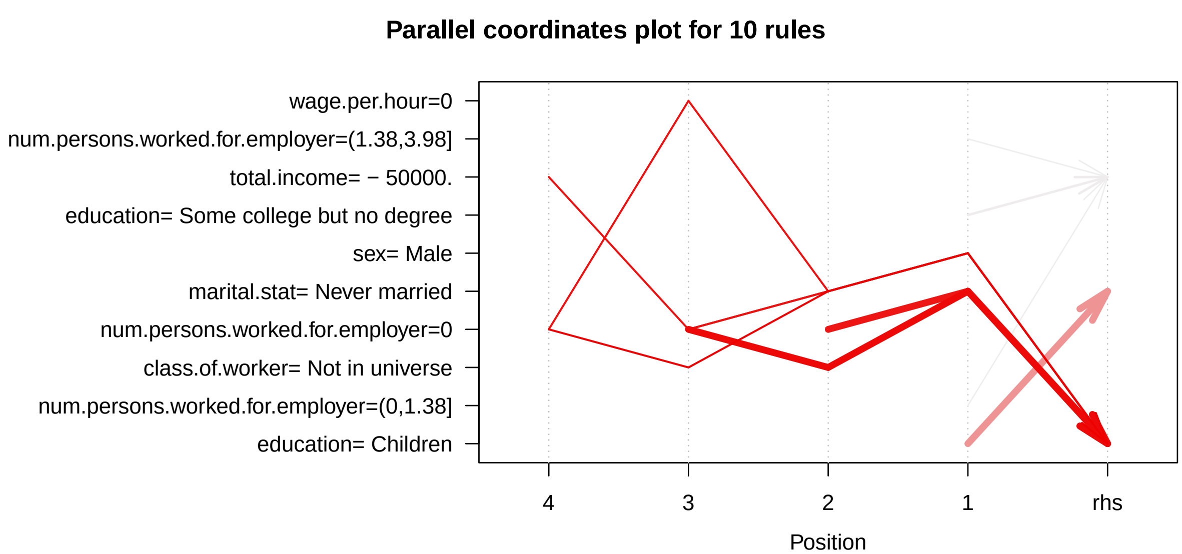
```
{weeks.worked.in.year=(40.5,52]} 0.0172 0.832
## {class.of.worker= State government} => {weeks.worked.in.year=(40.5,52]} 0.0172 0.8141
## {class.of.worker= Local
government,dividends.from.stocks=0,num.persons.worked.for.employer=(3.98,6]} =>
{weeks.worked.in.year=(40.5,52]} 0.0211 0.8087
## {class.of.worker= Local government,weeks.worked.in.year=(40.5,52]} =>
{num.persons.worked.for.employer=(3.98,6]} 0.0251 0.8204
## {class.of.worker= Local government,num.persons.worked.for.employer=(3.98,6]} =>
{weeks.worked.in.year=(40.5,52]} 0.0251 0.8048
## {class.of.worker= Local government} => {num.persons.worked.for.employer=(3.98,6]}
0.0311 0.7983
```

Al eliminar las reglas que terminan en valores muy frecuentes, obtenemos asociaciones que nos llaman la atención ya que afirman, por ejemplo, que si un individuo trabaja para el gobierno, no está desempleado y su sueldo es 0 por hora, entonces trabaja todo el año. Este tipo de asociaciones tienen un soporte de al menos el 0.15% con una confianza alta, sobre el 80%, con lo que podrían indicar que el atributo `wage.per.hour` tiene datos erróneos.

## Reglas de asociación generales

En esta sección construiremos reglas de asociación con ambos tipos de atributos. No podemos emplear todas las variables al mismo tiempo porque la complejidad del algoritmo es excesiva, sabemos que debemos tomar un soporte pequeño, con lo que se generarían demasiadas reglas. Seleccionamos un grupo de atributos interesante.

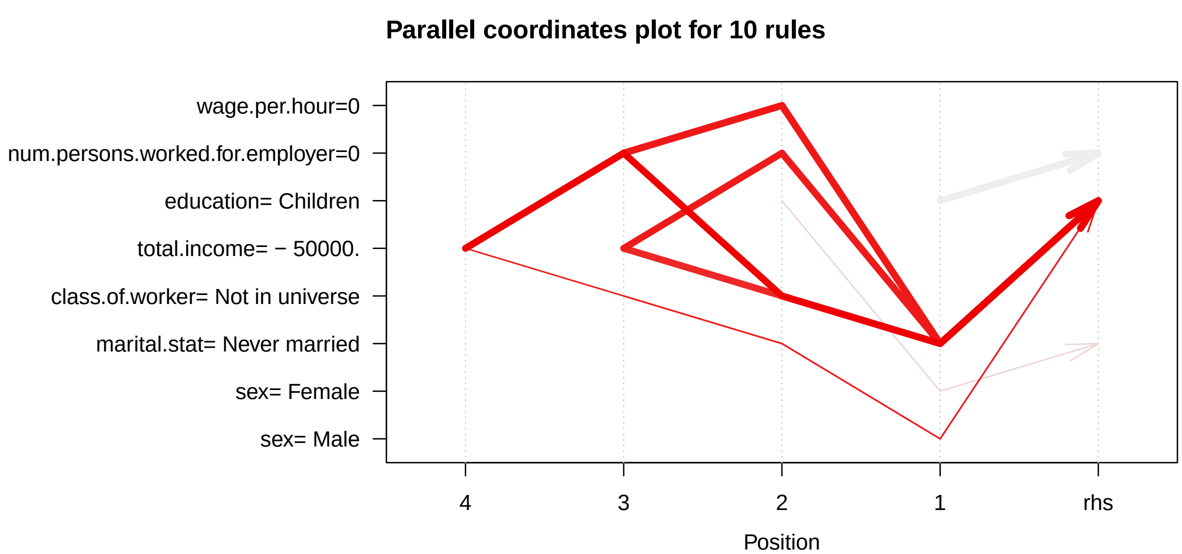
```r
reglas <- reglas_weclat(c("age", "sex", "marital.stat",
                          "class.of.worker", "education", "wage.per.hour",
                          "num.persons.worked.for.employer",
                          "total.income"), sop=0.1, 0.7)
reglas.sub <- subset(reglas, subset= !(rhs %oin% c(
  "class.of.worker= Not in universe", "wage.per.hour=0")))
paracoord_reglas(reglas.sub[1:10])
```



```r
print_reglas(reglas.sub[1:10])
```

```
## {num.persons.worked.for.employer=(0,1.38]} => {total.income= - 50000.} 0.1053 0.9094
## {num.persons.worked.for.employer=(1.38,3.98]} => {total.income= - 50000.} 0.1074
0.9116
## {sex= Male,marital.stat= Never married,num.persons.worked.for.employer=0} =>
{education= Children} 0.1207 0.8081
## {sex= Male,marital.stat= Never married,class.of.worker= Not in
universe,num.persons.worked.for.employer=0} => {education= Children} 0.1207 0.8359
## {sex= Male,marital.stat= Never married,num.persons.worked.for.employer=0,total.income=
- 50000.} => {education= Children} 0.1207 0.8089
## {sex= Male,marital.stat= Never
married,wage.per.hour=0,num.persons.worked.for.employer=0} => {education= Children}
0.1207 0.8101
## {education= Some college but no degree} => {total.income= - 50000.} 0.1305 0.9358
## {education= Children} => {marital.stat= Never married} 0.2376 0.9999
## {marital.stat= Never married,num.persons.worked.for.employer=0} => {education=
Children} 0.2376 0.795
## {marital.stat= Never married,class.of.worker= Not in
universe,num.persons.worked.for.employer=0} => {education= Children} 0.2376 0.8219
```

```r
plot(reglas.sub[11:20], method="paracoord", reorder=TRUE)
```




```r
print_reglas(reglas.sub[11:20])
```

```
## {sex= Female,education= Children} => {marital.stat= Never married} 0.117 0.9999
## {sex= Male,marital.stat= Never married,class.of.worker= Not in universe} =>
{education= Children} 0.1207 0.7917
## {sex= Male,marital.stat= Never married,class.of.worker= Not in universe,total.income=
- 50000.} => {education= Children} 0.1207 0.7928
## {marital.stat= Never married,class.of.worker= Not in
universe,num.persons.worked.for.employer=0,total.income= - 50000.} => {education=
```

```
Children} 0.2376 0.8225
## {marital.stat= Never married,num.persons.worked.for.employer=0,total.income= - 50000.}
=> {education= Children} 0.2376 0.7956
## {marital.stat= Never
married,wage.per.hour=0,num.persons.worked.for.employer=0,total.income= - 50000.} =>
{education= Children} 0.2376 0.7976
## {marital.stat= Never married,wage.per.hour=0,num.persons.worked.for.employer=0} =>
{education= Children} 0.2376 0.7971
## {marital.stat= Never married,class.of.worker= Not in universe} => {education=
Children} 0.2376 0.7803
## {marital.stat= Never married,class.of.worker= Not in universe,total.income= - 50000.}
=> {education= Children} 0.2376 0.7811
## {education= Children} => {num.persons.worked.for.employer=0} 0.2377 1
```

En el caso general, conseguimos encontrar asociaciones más informativas porque relacionan las variables sociales con las económicas.

# Bibliografía


[1] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, ACM SIGMOD Record. 22 (1993) 207–216. https://doi.org/10.1145/170036.170072.

[2] G. Subramanian, R data analysis projects : Build end to end analytics systems to get deeper insights from your data, Packt Publishing, Birmingham, UK, 2017.

[3] W. Wang, J. Yang, P.S. Yu, Efficient mining of weighted association rules (WAR), in: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 00, ACM Press, Boston, Massachusetts, USA, 2000. https://doi.org/10.1145/347090.347149.

[4] D. Dua, C. Graff, UCI machine learning repository, (2017). http://archive.ics.uci.edu/ml.