ANÁLISIS FACTORIAL PARA DATOS DE RURALIDAD CON R

Autor:

Luis Antonio Martínez Hernández

16 De Junio, 2022

Contents

Introducción

Durante este proyecto de análisis factorial, estaremos trabajando con datos de ruralidad, sin embargo, antes de ello, necesitamos comprender a qué nos referimos con ruralidad, puesto que su concepto y definición pueden llegar a crear confusiones. *La ruralidad es un concepto de moda. Diferentes publicaciones científicas y técnicas, responsables políticos y gremiales utilizan más que nunca este término, debido antes que nada a la emergencia de nuevas formas de vida y nuevas dinámicas de desarrollo en los espacios rurales, a tal punto que aparece en diversos ámbitos el concepto de "nueva ruralidad". Sin embargo, esta idea tan en boga mantiene un alto grado de ambigüedad, lo que genera aún mayor confusión.*

En líneas generales, la ruralidad tiene dos grandes acepciones y usos. La primera acepción, que es la más banal y ampliamente difundida, hace referencia a la ruralidad como "todos los hechos y fenómenos relativos que se suceden en áreas de baja densidad de población vinculada a la producción de bienes primarios o agropecuarios". Esta es una concepción estática que tiene un fuerte carácter demográfico y espacial.

La segunda acepción, que es la que consideramos pertinente para abordar en profundidad los procesos de desarrollo rural, define a la ruralidad como "la forma de relación que se establece entre la sociedad y los espacios rurales y a partir de la cual, se construye el sentido social de lo rural, la identidad y se moviliza el patrimonio de dichos espacios".

La ruralidad en tanto forma de relación entre el hombre y su espacio y forma de apropiación simbólica, valorización y aprovechamiento del patrimonio, constituye la dimensión social de los territorios rurales. Entendida de esta manera, la ruralidad tiene dos dimensiones que interesa analizar: a) una dimensión subjetiva vinculada a la identidad, y b) una dimensión instrumental ligada a las formas de valorización del patrimonio. Ambas dimensiones son concurrentes al proceso de apropiación y territorialización de los espacios rurales.

Teniendo esta visión más amplia de qué es ruralidad, podemos cambiar el uso de “ruralidad” por “zonas o espacios rurales”, así, procedemos a iniciar nuestro análisis factorial, con el fin de reducir las variables originales a un número menor de factores, los cuales nos puedan ayudar a entender mucho mejor aspectos de nuestros datos.

1. Datos

Aplicaremos un análisis de factores a datos rurales del año 2020, las variables se describen a continuación:

Factor	**Subgrupo**	**Variable**	**Nombre**
Social	Localidad	LocChi	Porcentaje de la población que habita en localidades chicas.
Social	Localidad	LocMed	Porcentaje de la población que habita en localidades medianas.
Social	Localidad	LocGra	Porcentaje de la población que habita en localidades grandes.
Social	Localidad	CamPob	Cambio de población (cambio porcentual de la población).
Social	Vivienda	InAgu	Porcentaje de agua entubada.
Social	Vivienda	InEle	Porcentaje de electricidad.
Social	Vivienda	InDre	Porcentaje de drenaje.
Social	Vivienda	MatViv	Materiales del techo de las viviendas (Porcentaje).
Económico	Laboral	PeaDes	Porcentaje de Población Económicamente Activa desocupada.
Económico	Laboral	PeaPrim	% de pobl. económicamente activa ocupada en el sect. primario.
Económico	Laboral	PeaSec	% de pobl. económicamente activa ocupada en el sect. secundario.
Económico	Laboral	PeaTer	% de pobl. económicamente activa ocupada en el sect. terciario.

Haremos uso de los tres métodos vistos en clase:

- **Método de máxima verosimilitud con y sin rotación.**
- **Método de factores principales con y sin rotación.**
- **Método de componentes principales con y sin rotación.**

Los métodos se aplicarán a cada subgrupo de datos: subgrupo Localidad, subgrupo Vivienda y subgrupo Laboral, en ese orden.

2. Subgrupo Localidad

Empezamos por separar el subgrupo Localidad de la base de datos.

```
localidad <- ruralidad[,5:8]
head(localidad,5)
```

```
 # A tibble: 5 x 4
  LocChi LocMed LocGra CamPob
   <dbl>  <dbl>  <dbl>  <dbl>
1   5.01   3.96   91.0   8.19
2  65.5   34.5     0    10.9
3  51.3   12.6    36.1   3.93
4  50.1   49.9     0     9.14
5  24.9   26.0    49.1   7.91
```

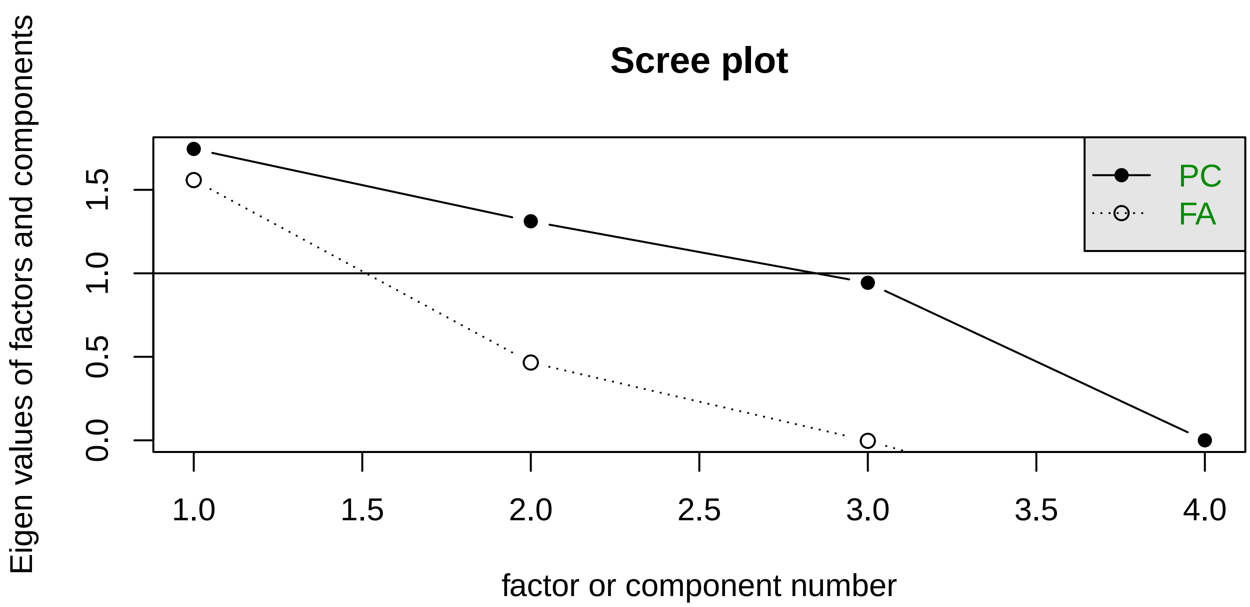
Calculamos la matriz de correlaciones muestral $\boldsymbol{R}$ para este subgrupo:

```
R_loc <- cor(localidad); R_loc
```

```
           LocChi      LocMed     LocGra      CamPob
LocChi  1.0000000 -0.60936235 -0.5731639 -0.14073607
LocMed -0.6093623  1.00000000 -0.3004635  0.04162117
LocGra -0.5731639 -0.30046346  1.0000000  0.12628083
CamPob -0.1407361  0.04162117  0.1262808  1.00000000
```

Ahora, determinamos el número de factores a extraer, solo nos interesarán los factores que tengan un eigenvalue mayor a 1.0 o eigenvalues superiores a los demás. Una manera de encontrar estos factores es utilizando una *screeplot*, como a continuación se muestra:

```
scree(R_loc)
```



La *screeplot* nos sugiere elegir entre 1 y 2 factores, de este modo, nos decantamos por elegir dos factores, sin embargo, para componentes principales es mejor usar 3 factores ya que tiene eigenvalores más grandes, como se puede ver en la gráfica.

2.1 Método de Máxima Verosimilitud Sin Rotación

Aplicamos el método de máxima verosimilitud a los datos del subgrupo Localidad, para ello nos apoyaremos de la librería `psych` usando la función `fa()` como a continuación se muestra:

```
library(psych)
# Aplicamos el método de máx. verosimilitud indicando fm="ml"
fit.m1 <- fa(R_loc, nfactors=2, n.iter=1, rotate="none", fm="ml")
fit.m1
```

```
 Factor Analysis using method =  ml
Call: fa(r = R_loc, nfactors = 2, n.iter = 1, rotate = "none", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
         ML1   ML2    h2     u2 com
LocChi -1.00 -0.03 0.997 0.0029 1.0
LocMed  0.63 -0.77 0.997 0.0035 1.9
LocGra  0.54  0.84 0.996 0.0036 1.7
CamPob  0.14  0.06 0.023 0.9770 1.4

                       ML1  ML2
SS loadings           1.71 1.30
Proportion Var        0.43 0.32
Cumulative Var        0.43 0.75
Proportion Explained  0.57 0.43
Cumulative Proportion 0.57 1.00

Mean item complexity =  1.5
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  6  and the objective function was  24.21
The degrees of freedom for the model are -1  and the objective function was  18.26

The root mean square of the residuals (RMSR) is  0
The df corrected root mean square of the residuals is  NA

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   ML1  ML2
Correlation of (regression) scores with factors   1.00 1.00
Multiple R square of scores with factors          1.00 1.00
Minimum correlation of possible factor scores     0.99 0.99
```

La primera parte de la salida del programa nos muestra los pesos para los dos factores y también nos muestra las comunalidades (`h2`) y varianzas específicas (`u2`).
Para el primer factor se obtiene un gran peso negativo para la variable *LocChi* (porcentaje de población en localidades chicas) y pesos positivos aunque no tan grandes para *LocMed* (porcentaje en localidades medianas), *LocGra* (porcentaje en localidades grandes) y *CamPob* (cambio porcentual de la población). Podemos decir, que el primer factor contrasta la característica esencial de un espacio rural (el de localidades chicas) con el de características que no son propias de un espacio rural, a excepción de *CamPob* que no influye en gran medida.
Para el segundo factor obtenemos pesos grandes en *LocMed* y *LocGra* a diferencia de *LocChi* y *CamPob* que tienen pesos muy bajos, de este modo, podemos decir que el segundo factor influye más en características contrarias a las de rural. En ambos casos, el cambio de población (*CamPob*) no es muy influyente, por lo que podemos decir que no está relacionado en gran medida con lo rural.

2.2 Método de Máxima Verosimilitud Con Rotación

Aplicamos el método de máxima verosimilitud junto **con el método de rotación varimax** a los datos del subgrupo Localidad como a continuación se muestra:

```
# Aplicamos el método de máx. verosimilitud indicando fm="ml" y rotacion varimax
fit.m1.rot <- fa(R_loc, nfactors=2, n.iter=1, rotate="varimax", fm="ml")
fit.m1.rot
```

```
 Factor Analysis using method =  ml
Call: fa(r = R_loc, nfactors = 2, n.iter = 1, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
         ML1   ML2    h2     u2 com
LocChi -0.80 -0.60 0.997 0.0029 1.9
LocMed  0.01  1.00 0.997 0.0035 1.0
LocGra  0.95 -0.31 0.996 0.0036 1.2
CamPob  0.15  0.04 0.023 0.9770 1.2

                       ML1  ML2
SS loadings           1.56 1.46
Proportion Var        0.39 0.36
Cumulative Var        0.39 0.75
Proportion Explained  0.52 0.48
Cumulative Proportion 0.52 1.00

Mean item complexity =  1.3
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  6  and the objective function was  24.21
The degrees of freedom for the model are -1  and the objective function was  18.26

The root mean square of the residuals (RMSR) is  0
The df corrected root mean square of the residuals is  NA

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 ML1  ML2
Correlation of (regression) scores with factors   1.00 1.00
Multiple R square of scores with factors          1.00 1.00
Minimum correlation of possible factor scores     0.99 0.99
```

Aplicando la rotación **varimax** vemos que los pesos cambian con respecto al modelo sin rotar. En el modelo **con rotación**, para el primer factor, las variables en las que más se influyen son Localidades grandes ($LocGra$) y Localidades chicas ($LocChi$), con signo contrario, es decir, que el factor contrasta una zona rural tomando como caracteristica las localidades chicas, con zonas no rurales las cuales se caracterizan por localidades grandes.

Para su segundo factor, sucede algo parecido, pero esta vez contrastanto las localidades chicas con las medianas, ya que son las variables con más peso y con signo contrario, podemos decir, que lo rural toma como exclusivas las localidades chicas, dejando a un lado las localidades de mayor tamaño.

En ambos factores, el cambio de población ($CamPob$) no parece ser relevante a la hora de determinar lo rural. Las interpretaciones que se dieron en el modelo **sin rotar** fueron parecidas, aunque la interpretación para el segundo factor cambió ya que los pesos de mayor tamaño fueron para $LocMed$ y $LocGra$ (en ese caso, el factor influye más en las características de no rural).

2.3 Método de Factores Principales Sin Rotación

Ahora, aplicaremos el análisis de factores usando el método de factores principales, para ello, realizaremos una estimación inicial de las comunalidades haciendo una regresión de cada variable sobre el resto de ellas y obteniendo la comunalidad de una variable mediante el coeficiente de determinación. Creamos una función que nos calcule los pesos factoriales de esta forma:

```r
# Creamos la función para los pesos
fact.r2 <- function(data, nfact, iter){
  R <- cor(data); h2j <- c()
  for (i in 1:ncol(data)) {
    formula <- paste(names(data)[i], "~.", sep=" ")
    h2j <- c(h2j, summary(lm(formula, data))$r.squared)
  }
  for (i in 1:iter) {
    Psi <- diag(1-h2j); QQt <- R-Psi; E <- svd(QQt)
    Q <- (E$u%*%diag(sqrt(E$d)))[,1:nfact]
    if (i < iter) {if (nfact==1) {h2j <- Q**2} else {h2j <- rowSums((Q**2))}}
  }
  return(list(Q_hat=Q,h2_hat=h2j,Var_esp=diag(Psi),lambda=E$d,Psi_hat=Psi))
}
# Aplicamos la función para el subgrupo Localidad usando 5 iteraciones
fit.m2 <- fact.r2(localidad, 2, 5)
# Función para mostrar resultados
table1<- function(model, data, k,p){lambda=model$lambda;Q_hat<-data.frame(q=model$Q_hat)
  t1 <- cbind(round(Q_hat,2),round(cbind(Communalities=model$h2_hat,
                                         Specific_Var=model$Var_esp),4))
  t1 <- as.data.frame(cbind(Variables=names(data), t1))
  t2 <- as.data.frame(rbind(Eigenvalues=lambda[1:k], Proportion_Var=lambda[1:k]/p,
                            Cumulative_Var = cumsum(lambda[1:k]/p)))
  names(t2) <-  colnames(Q_hat); return(list(t1 = t1, t2=t2))
}
# Cargas factoriales estimadas, comunalidades y varianzas específicas
table1(fit.m2, localidad, k=2, p=4)
```

```
 $t1
  Variables   q.1   q.2 Communalities Specific_Var
1    LocChi -1.00 -0.01        1.0000       0.0000
2    LocMed  0.62 -0.78        1.0000       0.0000
3    LocGra  0.56  0.83        1.0000       0.0000
4    CamPob  0.14  0.06        0.0229       0.9771

$t2
                     q.1       q.2
Eigenvalues    1.7197224 1.3031828
Proportion_Var 0.4299306 0.3257957
Cumulative_Var 0.4299306 0.7557263
```

Obtenemos resultados aproximandamente iguales que con el método de máxima verosimilitud **sin rotación**, y por ende, las mismas interpretaciones; hemos de notar que la proporción acumulada de la varianza es de 75% lo cual es un proporción aceptable tomando en cuenta que solo son 2 factores los que se utilizaron. Además, las comunalidades son iguales a 1 en las tres primeras variables, que, como ya mencionamos, toman un papel importante a la hora de separar lo rural y lo no rural, a excepción de la última variable *CamPob*, la cual no es tan determinante como las demás.

2.4 Método de Factores Principales Con Rotación

Ahora, aplicaremos el método de factores principales junto **con el método de rotación varimax** usando **2 factores**, nos apoyaremos de la librería `psych`:

```r
# Aplicamos el método de factores princ. indicando fm="paf" y rotacion varimax
fit.m2.rot <- fa(R_loc, nfactors=2, n.iter=1, rotate="varimax", fm="paf")
fit.m2.rot
```

```
 Factor Analysis using method =  minres
Call: fa(r = R_loc, nfactors = 2, n.iter = 1, rotate = "varimax", fm = "paf")
Standardized loadings (pattern matrix) based upon correlation matrix
         MR1   MR2    h2     u2 com
LocChi -0.80 -0.60 0.997 0.0029 1.9
LocMed  0.01  1.00 0.997 0.0035 1.0
LocGra  0.95 -0.31 0.996 0.0036 1.2
CamPob  0.15  0.04 0.023 0.9770 1.2

                       MR1  MR2
SS loadings           1.56 1.46
Proportion Var        0.39 0.36
Cumulative Var        0.39 0.75
Proportion Explained  0.52 0.48
Cumulative Proportion 0.52 1.00

Mean item complexity =  1.3
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  6  and the objective function was  24.21
The degrees of freedom for the model are -1  and the objective function was  18.26

The root mean square of the residuals (RMSR) is  0
The df corrected root mean square of the residuals is  NA

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   MR1  MR2
Correlation of (regression) scores with factors   1.00 1.00
Multiple R square of scores with factors          1.00 1.00
Minimum correlation of possible factor scores     0.99 0.99
```

Notemos que los resultados son iguales a los obtenidos con el método de verosimilitud **con rotación**, por lo tanto, mantenemos las interpretaciones. Además, las comunalidades y la varianza acumulada se mantienen sin cambios en los modelos **sin rotar** y **con rotación**.

2.5 Método de Componentes Principales Sin Rotación

Por último, aplicaremos el método de componentes principales **sin rotación**. Sin embargo, a diferencia de los modelos anteriores, esta vez usaremos **3 factores** en lugar de 2. Creamos una función para realizar este método y lo aplicamos:

```r
# Localidad Componentes Principales
fact.comp <- function(R, k){
  Eigen <- svd(R)
  Q_hat <- ((Eigen$u)%*%diag(sqrt(c(Eigen$d))))[,1:k]
  if (k==1) {h2_hat <- Q_hat**2} else {h2_hat <- rowSums((Q_hat**2))}
  Psi_hat <- diag(1-h2_hat)
  return(list(Q_hat=Q_hat,h2_hat=h2_hat,Var_esp=diag(Psi_hat),
             lambda=Eigen$d,Psi_hat=Psi_hat))
}
# Aplicamos la función para el subgrupo Localidad usando la función creada
fit.m3 <- fact.comp(R_loc, 3)
# Cargas factoriales estimadas, comunalidades y varianzas específicas
table1(fit.m3, localidad, k=3, p=4)
```

```
 $t1
  Variables   q.1   q.2   q.3 Communalities Specific_Var
1    LocChi -0.98  0.04  0.18             1            0
2    LocMed  0.59 -0.81  0.01             1            0
3    LocGra  0.57  0.79 -0.22             1            0
4    CamPob  0.32  0.20  0.93             1            0

$t2
                    q.1       q.2       q.3
Eigenvalues    1.7448941 1.3116719 0.9434340
Proportion_Var 0.4362235 0.3279180 0.2358585
Cumulative_Var 0.4362235 0.7641415 1.0000000
```

Para Componentes Principales, usando 2 factores obtuvimos resultados parecidos a los otros métodos **sin rotar**, sin embargo, usando 3 factores, la variable *CamPob* toma muchísima más importancia que con los otros métodos (ya que también se probó usar 3 factores), y esto es muy notable en las comunalidades, las cuales son iguales a 1 en las 4 variables.

Los dos primeros factores siguen teniendo una interpretación similar a la que se hizo en el inicio, sin embargo, para el tercer factor, todo el peso cae en la variable de cambio de población (*CamPob*), dejando de ser relevantes las demás variables, dándonos a entender que el tercer factor solo tomará en cuenta a esta variable. Además, la proporción acumulada de la varianza llega a ser exactamente 1, por lo que para el método de componentes principales, usar 3 factores es mucho mejor que usar solo 2 como en los anteriores modelos.

2.6 Método de Componentes Principales Con Rotación

Por último, aplicaremos el método de componentes principales **con rotación varimax** y **3 factores**.

```r
# Aplicamos el método de componentes principales y rotacion varimax con 3 factores
fit.m3.rot <- principal(R_loc, nfactors=3, rotate="varimax")
fit.m3.rot
```

```
 Principal Components Analysis
Call: principal(r = R_loc, nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
         RC1   RC2   RC3 h2       u2 com
LocChi -0.74 -0.67 -0.07  1  5.8e-12 2.0
LocMed  0.98 -0.18  0.02  1  4.3e-12 1.1
LocGra -0.12  0.99  0.07  1  4.0e-12 1.0
CamPob  0.03  0.07  1.00  1 -3.6e-15 1.0

                       RC1  RC2  RC3
SS loadings           1.53 1.46 1.00
Proportion Var        0.38 0.37 0.25
Cumulative Var        0.38 0.75 1.00
Proportion Explained  0.38 0.37 0.25
Cumulative Proportion 0.38 0.75 1.00

Mean item complexity =  1.3
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0

Fit based upon off diagonal values = 1
```

El primer factor contrasta las localidades chicas ($LocChi$) (que son comúnes en una zona rural) con localidades medianas ($LocMed$), ya que son las de mayor peso y signo contrario. El segundo factor, tiene más influencia en las localidades chicas y las grandes, y como tienen signo contrario podemos decir, que se contrastan para diferenciar un zona rural y una no rural, esto pasó de manera similar en los dos primeros factores del modelo **sin rotar**, donde la diferencia principal se tomaba de las localidades chicas contrastando con las demás variables. El tercer factor tiene una interpretación igual a la que se hizo en el modelo **sin rotar**, es decir, el factor describe el cambio de población ($CamPob$) ya que es el de mayor peso el cual no había sido relevante en los primeros dos factores.

3. Subgrupo Vivienda

Separamos el subgrupo Vivienda de la base de datos.

```
vivienda <- ruralidad[,9:12]
head(vivienda, 5)
```

```
  # A tibble: 5 x 4
    inagu inelec indre MatViv
    <dbl>  <dbl> <dbl>  <dbl>
1   99.6   99.9  99.7   99.3
2   99.1   99.5  96.8   91.7
3   99.2   99.5  99.5   99.6
4   99.3   99.4  99.1   95.6
5   99.1   99.6  99.2   98.9
```

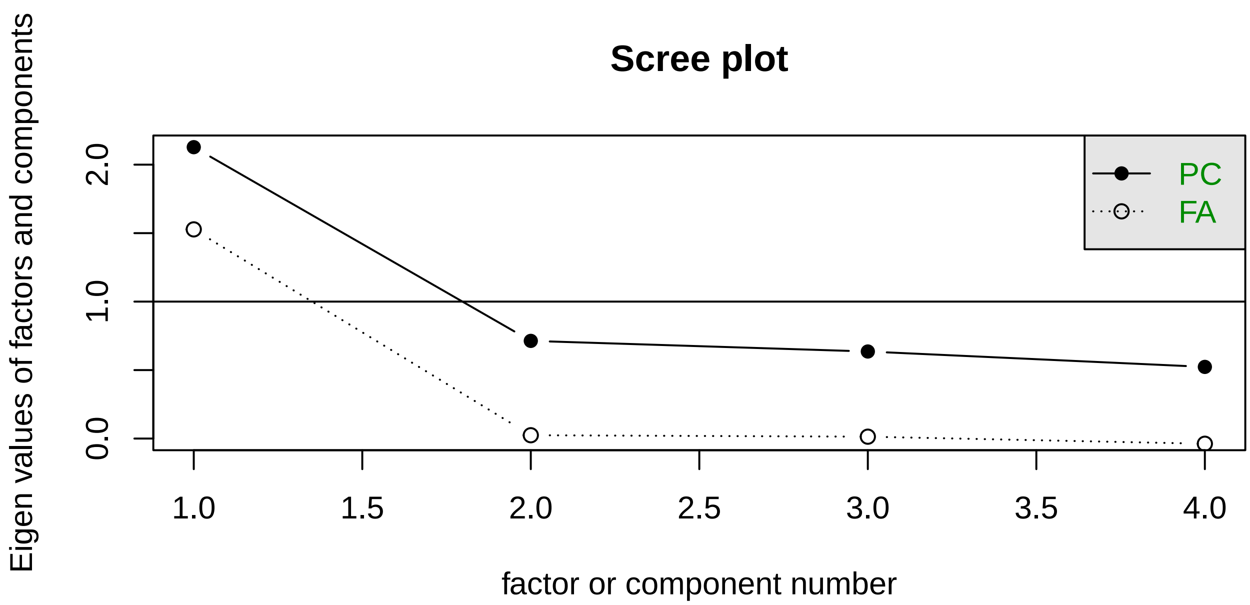
Calculamos la matriz de correlaciones muestral $\boldsymbol{R}$ para este subgrupo:

```
R_viv <- cor(vivienda); R_viv
```

```
           inagu    inelec     indre    MatViv
inagu  1.0000000 0.3437653 0.4133517 0.2869626
inelec 0.3437653 1.0000000 0.4295231 0.3462536
indre  0.4133517 0.4295231 1.0000000 0.4257194
MatViv 0.2869626 0.3462536 0.4257194 1.0000000
```

Ahora, determinamos el número de factores a extraer utilizando una *screeplot*:

```
scree(R_viv)
```



La *screeplot* nos sugiere elegir 1 factor, sin embargo, para componentes principales es mejor usar 3 factores ya que sus eigenvalores son un poco más grandes, como se puede ver en la gráfica.

3.1 Método de Máxima Verosimilitud Sin Rotación

Aplicamos el método de máxima verosimilitud a los datos del subgrupo Vivienda, para ello nos apoyaremos de la librería `psych` usando la función `fa()` como a continuación se muestra:

```
# Aplicamos el método de máx. verosimilitud indicando fm="ml" y solo 1 factor
fit.m1 <- fa(R_viv, nfactors=1, n.iter=1, rotate="none", fm="ml")
fit.m1
```

```
 Factor Analysis using method =  ml
Call: fa(r = R_viv, nfactors = 1, n.iter = 1, rotate = "none", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
        ML1   h2   u2 com
inagu  0.55 0.31 0.69   1
inelec 0.60 0.36 0.64   1
indre  0.74 0.55 0.45   1
MatViv 0.57 0.32 0.68   1

                ML1
SS loadings    1.53
Proportion Var 0.38

Mean item complexity =  1
Test of the hypothesis that 1 factor is sufficient.

The degrees of freedom for the null model are  6  and the objective function was  0.68
The degrees of freedom for the model are 2  and the objective function was  0

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.02

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 ML1
Correlation of (regression) scores with factors   0.85
Multiple R square of scores with factors          0.73
Minimum correlation of possible factor scores     0.45
```

La primera parte de la salida del programa nos muestra los pesos para el único factor y también nos muestra las comunalidades (`h2`) y varianzas específicas (`u2`).

Obtenemos pesos positivos para cada una de las variables, *inagu* (porcentaje de agua entubada), *inelec* (porcentaje de electricidad), *indre* (porcentaje de drenaje) y $MatViv$ (materiales del techo de las viviendas), los pesos son muy parecidos en cada variable, aunque, es más sobresaliente para el porcentaje de drenaje (*indre*), es decir, el factor equilibra las variables, inclinándose un poco más al drenaje para caracterizar un espacio no rural, puesto que en estas zonas es más común ver este tipo de infraestructura que en zonas rurales.

3.2 Método de Máxima Verosimilitud Con Rotación

Nuevamente aplicamos el método de verosimilitud, esta vez con **con rotación varimax**:

```
library(psych)
# Aplicamos el método de máx. verosimilitud indicando fm="ml" y rotacion varimax
fit.m1.rot <- fa(R_viv, nfactors=1, n.iter=1, rotate="varimax", fm="ml")
fit.m1.rot
```

```
 Factor Analysis using method =  ml
Call: fa(r = R_viv, nfactors = 1, n.iter = 1, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
        ML1   h2   u2 com
inagu  0.55 0.31 0.69   1
inelec 0.60 0.36 0.64   1
indre  0.74 0.55 0.45   1
MatViv 0.57 0.32 0.68   1

                ML1
SS loadings    1.53
Proportion Var 0.38

Mean item complexity =  1
Test of the hypothesis that 1 factor is sufficient.

The degrees of freedom for the null model are  6  and the objective function was  0.68
The degrees of freedom for the model are 2  and the objective function was  0

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.02

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   ML1
Correlation of (regression) scores with factors   0.85
Multiple R square of scores with factors          0.73
Minimum correlation of possible factor scores     0.45
```

Vemos que los resultados son los mismos para el modelo con rotación y el modelo sin rotar. Esto se debe a que solo se ocupó un solo factor. De este modo, las interpretaciones son las mismas.

3.3 Método de Factores Principales Sin Rotación

Ahora, aplicaremos el análisis de factores usando el método de factores principales sin rotar.

```r
# Aplicamos la función para el subgrupo Vivienda usando la función creada
# en el problema de Localidad, con 1 factor y 10 iteraciones
fit.m2 <- fact.r2(vivienda, 1, 10)
# Cargas factoriales estimadas, comunalidades y varianzas específicas
table1(fit.m2, vivienda, k=1, p=4)
```

```
 $t1
  Variables     q Communalities Specific_Var
1     inagu -0.55        0.3029       0.6971
2    inelec -0.60        0.3628       0.6372
3     indre -0.74        0.5448       0.4552
4    MatViv -0.56        0.3163       0.6837

$t2
                       q
Eigenvalues    1.5271034
Proportion_Var 0.3817759
Cumulative_Var 0.3817759
```

Los resultados para este método son aproximadamente iguales a los obtenidos con el método de verosimilud con y sin rotación, aunque el signo para los pesos es negativo en todas las variables, la interpretación sigue siendo la misma, teniendo a la variable *indre* como la más influyente. Además, en las comunalidades también se puede ver que el porcentaje de drenaje tiene un valor mayor que el de las demás. La proporción de la varianza que explica este factor es de 38.2% que, pese a que no es mucho, sí es significativa que si se usa más de un factor (ya que se probó usar más de dos factores pero los resultados no cambiaron en gran medida y nos decantamos por usar solo un factor).

3.4 Método de Factores Principales Con Rotación

Ahora aplicamos el método con la rotación varimax:

```r
# Aplicamos el método de factores princ. indicando fm="paf" y rotacion varimax
fit.m2.rot <- fa(R_viv, nfactors=1, n.iter=1, rotate="varimax", fm="paf")
fit.m2.rot
```

```
 Factor Analysis using method =  minres
Call: fa(r = R_viv, nfactors = 1, n.iter = 1, rotate = "varimax", fm = "paf")
Standardized loadings (pattern matrix) based upon correlation matrix
        MR1   h2   u2 com
inagu  0.55 0.30 0.70   1
inelec 0.60 0.36 0.64   1
indre  0.74 0.55 0.45   1
MatViv 0.56 0.32 0.68   1

                MR1
SS loadings    1.53
Proportion Var 0.38

Mean item complexity =  1
Test of the hypothesis that 1 factor is sufficient.

The degrees of freedom for the null model are  6  and the objective function was  0.68
The degrees of freedom for the model are 2  and the objective function was  0

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.02

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   MR1
Correlation of (regression) scores with factors   0.85
Multiple R square of scores with factors          0.73
Minimum correlation of possible factor scores     0.46
```

Obtenemos los mismos resultados que en los modelos anteriores y las mismas interpretaciones.

3.5 Método de Componentes Principales Sin Rotación

Ahora, aplicaremos el método de componentes principales usando la función definida en el problema anterior de Localidad, además, en este caso se usarán tres factores a diferencia de los métodos anteriores:

```
# Aplicamos la función para el subgrupo Vivienda usando la función creada
# Usando 3 factores
fit.m3 <- fact.comp(R_viv, 3)

# Cargas factoriales estimadas, comunalidades y varianzas específicas
table1(fit.m3, vivienda, k=3, p=4)
```

```
 $t1
  Variables   q.1   q.2   q.3 Communalities Specific_Var
1     inagu -0.69  0.61 -0.31        0.9471       0.0529
2    inelec -0.73  0.00  0.66        0.9660       0.0340
3     indre -0.79 -0.02 -0.06        0.6350       0.3650
4    MatViv -0.70 -0.58 -0.32        0.9287       0.0713

$t2
                      q.1       q.2       q.3
Eigenvalues     2.1277458 0.7132719 0.6357131
Proportion_Var 0.5319364 0.1783180 0.1589283
Cumulative_Var 0.5319364 0.7102544 0.8691827
```

Para componentes principales sin rotar, se usaron tres factores, en donde, para el primer factor obtenemos pesos con un patrón similar a los de los métodos de verosimilitud y de factores principales, de este modo, podemos mantener la interpretación para este factor, en el cual se le da mucha más prioridad a el drenaje a comparación de las demás variables que se mantienen en un rango parecido.

Para el segundo factor, obtenemos pesos positivos para el porcentaje de agua ($inagu$) y electricidad ($inelec$), y pesos negativos para el porcentaje de drenaje ($indre$) y materiales del techo ($MatViv$). Notemos que en este segundo factor, los pesos de mayor magnitud son para el porcentaje de agua ($inagu$) y el de materiales del techo ($MatViv$), con signos contrarios.

Entonces, podemos interpretar que para una zona rural, el porcentaje de agua es esencial y determinante, puesto que en México varias zonas rurales se ubican en el sur, donde el agua es abundante, y esto, es contrastado con los materiales de vivienda, ya que, por lo general, en las zonas rurales no se cuentan con techos de buena calidad e incluso se ocupan techos con materiales improvisados.

Para el tercer factor, vemos que la variable de mayor influencia es de la electricidad ($inelec$), esto nos da un indicio de que el factor divide zonas rurales y no rurales de acuerdo a la electricidad, ya que por lo general, en lo zonas rurales se carece de este servicio, a diferencia de las no rurales. La proporción acumulada de la varianza llega hasta un 87% que es mucho más de lo que obtuvimos con los modelos anteriores.

3.6 Método de Componentes Principales Con Rotación

El método con rotación arroja los siguientes resultados:

```r
# Aplicamos el método de componentes principales y rotacion varimax con 3 factores
fit.m3.rot <- principal(R_viv, nfactors=3, rotate="varimax")
fit.m3.rot
```

```
 Principal Components Analysis
Call: principal(r = R_viv, nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
        RC1  RC2  RC3   h2    u2 com
inagu  0.12 0.96 0.14 0.95 0.053 1.1
inelec 0.16 0.15 0.96 0.97 0.034 1.1
indre  0.52 0.45 0.40 0.64 0.365 2.9
MatViv 0.95 0.10 0.13 0.93 0.071 1.1

                       RC1  RC2  RC3
SS loadings           1.21 1.15 1.11
Proportion Var        0.30 0.29 0.28
Cumulative Var        0.30 0.59 0.87
Proportion Explained  0.35 0.33 0.32
Cumulative Proportion 0.35 0.68 1.00

Mean item complexity =  1.5
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.1

Fit based upon off diagonal values = 0.92
```

Para el primer factor del modelo **con rotación** los pesos más grandes son para *indre* y $MatViv$ (que corresponden al drenaje y al material del techo de la vivienda, respectivamente), entonces, se puede decir que el primer factor describe las características de una zona no rural, donde el drenaje y el material de la vivienda son de mejor calidad que en las zonas rurales (esta interpretación es parecida a la del modelo sin rotación donde se le dió más prioridad al drenaje). El segundo factor es claramente influenciado por la primer variable, la del agua (similar al modelo sin rotar), ya que los pesos para las demás variables son mucho más bajas, entonces, podemos interpretar que para una zona rural, el porcentaje de agua es esencial y determinante. Para el tercer factor, el peso de mayor magnitud es el de la electricidad (*inelec*) y es lo mismo que pasa en el modelo sin rotar, de este modo, el tercer factor divide zonas rurales y no rurales de acuerdo a la electricidad.

4. Subgrupo Laboral

Empezamos por separar el subgrupo Laboral de la base de datos.

```
laboral <- ruralidad[,13:16]
head(laboral,5)
```

```
 # A tibble: 5 x 4
  PeaDes PeaPrim PeaSec PeaTer
   <dbl>   <dbl>  <dbl>  <dbl>
1   2.84    1.62   27.6   70.8
2   4.36   24.6    39.8   35.6
3   6.47   21.0    31.4   47.6
4   3.03   19.3    47.2   33.5
5   1.62    3.82   33.8   62.3
```

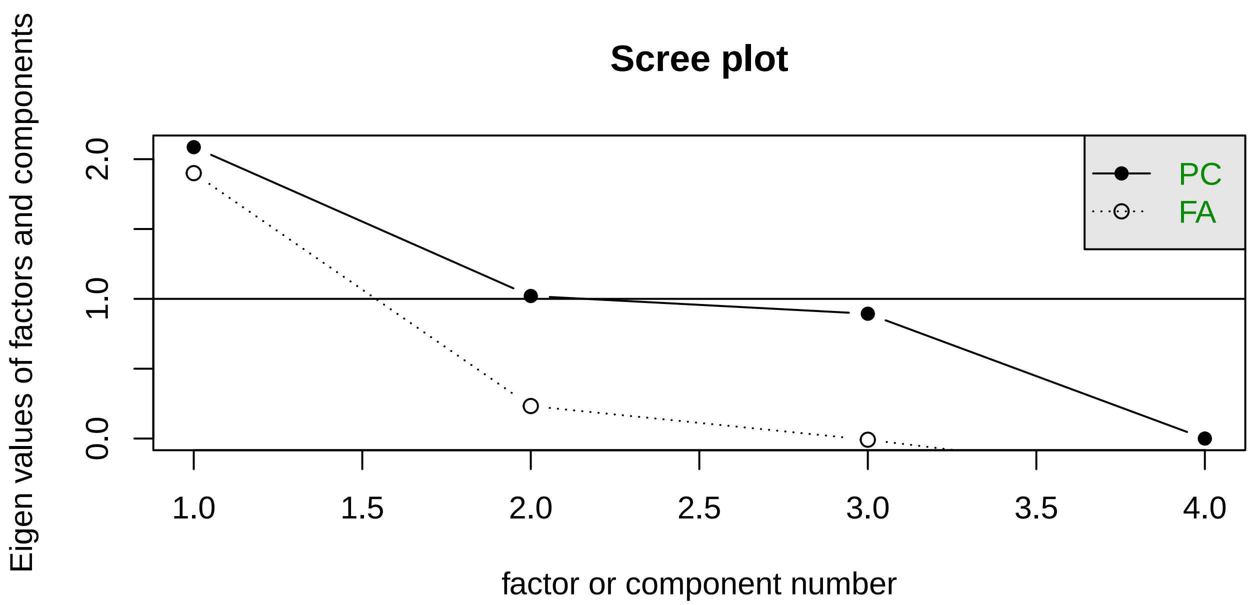
Calculamos la matriz de correlaciones muestral $\boldsymbol{R}$ para este subgrupo:

```
R_lab <- cor(laboral); R_lab
```

```
              PeaDes      PeaPrim      PeaSec      PeaTer
PeaDes   1.000000000  0.008967915  0.03417976 -0.03245140
PeaPrim  0.008967915  1.000000000 -0.58563238 -0.86116060
PeaSec   0.034179763 -0.585632380  1.00000000  0.09228062
PeaTer  -0.032451398 -0.861160602  0.09228062  1.00000000
```

Ahora, determinamos el número de factores a extraer utilizando una *screeplot*:

```
scree(R_lab)
```



La *screeplot* nos sugiere elegir entre 1 y 2 factores, de este modo, nos decantamos por elegir dos factores, sin embargo, para componentes principales es mejor usar 3 factores ya que tiene eigenvalores más grandes, como se puede ver en la gráfica.

4.1 Método de Máxima Verosimilitud Sin Rotación

Aplicamos el método de máxima verosimilitud sin rotar a los datos del subgrupo Laboral como a continuación se muestra:

```
# Aplicamos el método de máx. verosimilitud indicando fm="ml" y 2 factores
fit.m1 <- fa(R_lab, nfactors=2, n.iter=1, rotate="none", fm="ml")
fit.m1
```

```
 Factor Analysis using method =  ml
Call: fa(r = R_lab, nfactors = 2, n.iter = 1, rotate = "none", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
          ML1   ML2     h2     u2 com
PeaDes  -0.01  0.05 0.0025 0.9975 1.0
PeaPrim -1.00  0.03 0.9976 0.0024 1.0
PeaSec   0.61  0.79 0.9957 0.0043 1.9
PeaTer   0.84 -0.53 0.9967 0.0033 1.7

                       ML1  ML2
SS loadings           2.08 0.91
Proportion Var        0.52 0.23
Cumulative Var        0.52 0.75
Proportion Explained  0.70 0.30
Cumulative Proportion 0.70 1.00

Mean item complexity =  1.4
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  6  and the objective function was  21.3
The degrees of freedom for the model are -1  and the objective function was  15.13

The root mean square of the residuals (RMSR) is  0
The df corrected root mean square of the residuals is  NA

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                  ML1  ML2
Correlation of (regression) scores with factors     1 1.00
Multiple R square of scores with factors            1 0.99
Minimum correlation of possible factor scores       1 0.99
```

En el primer factor, las dos primeras variables *PeaDes* (% población economic. activa desocupada) y *PeaPrim* (% población económic. activa sector primario) tienen pesos negativos aunque *PeaDes* tiene un peso no significativo; las dos últimas variables *PeaSec* (% población económic. activa sector secundario) y *PeaTer* (% población económic. activa sector terciario) tienen pesos positivos y significativos. Entonces, podemos interpretar el factor 1 como el contraste entre la población ocupada en el sector primario (*PeaPrim*) y la población ocupada en el sector secundario y terciario (*PeaSec* y *PeaTer*), lo cual es lógico ya que en una zona rural es común que las personas se dediquen al sector primario en comparación con zonas no rurales donde es más frecuente que se dediquen al sector secundario o terciario.

Para el segundo factor tenemos que las variables de mayor peso son las dos últimas (*PeaSec* y *PeaTer*) con signos contrarios, y esto puede significar un contraste entre ambos, es decir, en zonas no rurales, las diferencias pueden estar marcadas por el porcentaje de la población que se dedique al sector secundario o el terciario, pues puede haber lugares donde el desarrollo del sector secundario es mayor que el sector terciario y viceversa.

4.2 Método de Máxima Verosimilitud Con Rotación

Luego, el método con rotación:

```r
library(psych)
# Aplicamos el método de máx. verosimilitud indicando fm="ml" y rotacion varimax
fit.m1.rot <- fa(R_lab, nfactors=2, n.iter=1, rotate="varimax", fm="ml")
fit.m1.rot
```

```
 Factor Analysis using method =  ml
Call: fa(r = R_lab, nfactors = 2, n.iter = 1, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
          ML1   ML2     h2     u2 com
PeaDes  -0.02  0.04 0.0025 0.9975 1.4
PeaPrim -0.96 -0.26 0.9976 0.0024 1.1
PeaSec   0.35  0.93 0.9957 0.0043 1.3
PeaTer   0.96 -0.26 0.9967 0.0033 1.1

                       ML1  ML2
SS loadings           1.98 1.01
Proportion Var        0.50 0.25
Cumulative Var        0.50 0.75
Proportion Explained  0.66 0.34
Cumulative Proportion 0.66 1.00

Mean item complexity =  1.3
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  6  and the objective function was  21.3
The degrees of freedom for the model are -1  and the objective function was  15.13

The root mean square of the residuals (RMSR) is  0
The df corrected root mean square of the residuals is  NA

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   ML1  ML2
Correlation of (regression) scores with factors   1.00 1.00
Multiple R square of scores with factors          1.00 0.99
Minimum correlation of possible factor scores     0.99 0.99
```

Para el modelo **con rotación**, el primer factor obtiene pesos grandes y con signos contrarios para las variables $PeaPrim$ y $PeaTer$ (que representan a la población activa en el sector primario y terciario, respectivamente), similar a lo que pasa en el modelo sin rotar, entonces, podemos interpretar el factor 1 como el contraste entre la población ocupada en el sector primario y la población ocupada en el sector terciario, separando una zona rural, donde las personas se dedican al sector primario, de una zona no rural, donde se dedican al sector terciario. Para el segundo factor, tenemos que la variable con más peso es la del sector secundario, de este modo, el factor 2 se centra en caracterizar solo a la población no rural que se dedica al sector secundario.

4.3 Método de Factores Principales Sin Rotación

Aplicaremos el método de factores principales primero sin rotar:

```r
# Aplicamos la función para el subgrupo Laboral usando la función creada
# en el problema de Localidad, con 2 factores y 10 iteraciones
fit.m2 <- fact.r2(laboral, 2, 10)
# Cargas factoriales estimadas, comunalidades y varianzas específicas
table1(fit.m2, laboral, k=2, p=4)
```

```
 $t1
  Variables   q.1   q.2 Communalities Specific_Var
1    PeaDes -0.01  0.05        0.0024       0.9976
2   PeaPrim -1.00  0.03        1.0000       0.0000
3    PeaSec  0.61  0.79        1.0000       0.0000
4    PeaTer  0.84 -0.53        1.0000       0.0000

$t2
                      q.1       q.2
Eigenvalues    2.0857852 0.9166613
Proportion_Var 0.5214463 0.2291653
Cumulative_Var 0.5214463 0.7506116
```

Los resultados para este método son aproximadamente iguales a los obtenidos con el método de verosimilud **sin rotación**, por ende, podemos mantener las interpretaciones ya hechas. Además, en ambos factores, el porcentaje de población economicamente activa desocupada ($PeaDes$) no es significativo, por lo cual, no es determinante para definir un espacio rural, esto es muy notable en las comunalidades de las variables, en donde se tienen valores grandes excepto para la primera. Hemos de notar que la proporción acumulada de la varianza es de 75% lo cual es un proporción aceptable tomando en cuenta que solo son 2 factores los que se utilizaron.

4.4 Método de Factores Principales Con Rotación

El método de factores principales con rotación:

```r
# Aplicamos el método de factores princ. indicando fm="paf" y rotacion varimax
fit.m2.rot <- fa(R_lab, nfactors=2, n.iter=1, rotate="varimax", fm="paf")
fit.m2.rot
```

```
 Factor Analysis using method =  minres
Call: fa(r = R_lab, nfactors = 2, n.iter = 1, rotate = "varimax", fm = "paf")
Standardized loadings (pattern matrix) based upon correlation matrix
          MR1   MR2     h2     u2 com
PeaDes  -0.02  0.04 0.0025 0.9975 1.4
PeaPrim -0.96 -0.26 0.9976 0.0024 1.1
PeaSec   0.35  0.93 0.9957 0.0043 1.3
PeaTer   0.96 -0.26 0.9967 0.0033 1.1

                       MR1  MR2
SS loadings           1.98 1.01
Proportion Var        0.50 0.25
Cumulative Var        0.50 0.75
Proportion Explained  0.66 0.34
Cumulative Proportion 0.66 1.00

Mean item complexity =  1.3
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  6  and the objective function was  21.3
The degrees of freedom for the model are -1  and the objective function was  15.13

The root mean square of the residuals (RMSR) is  0
The df corrected root mean square of the residuals is  NA

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   MR1  MR2
Correlation of (regression) scores with factors   1.00 1.00
Multiple R square of scores with factors          1.00 0.99
Minimum correlation of possible factor scores     0.99 0.99
```

Los resultados para este método son aproximadamente iguales a los obtenidos con el método de verosimilud **con rotación**, por ende, podemos mantener las interpretaciones ya hechas.

4.5 Método de Componentes Principales Sin Rotación

Aplicaremos el método de componentes principales sin rotar, además, en este caso se usarán tres factores a diferencia de los modelos anteriores:

```
# Aplicamos la función para el subgrupo Laboral usando la función creada
# Usando tres factores
fit.m3 <- fact.comp(R_lab, 3)
# Cargas factoriales estimadas, comunalidades y varianzas específicas
table1(fit.m3, laboral, k=3, p=4)
```

```
 $t1
  Variables   q.1   q.2   q.3 Communalities Specific_Var
1    PeaDes  0.01 -0.92  0.38             1            0
2   PeaPrim  1.00 -0.01 -0.03             1            0
3    PeaSec -0.61 -0.34 -0.71             1            0
4    PeaTer -0.84  0.22  0.49             1            0

$t2
                    q.1       q.2       q.3
Eigenvalues    2.085836 1.0204795 0.8936844
Proportion_Var 0.521459 0.2551199 0.2234211
Cumulative_Var 0.521459 0.7765789 1.0000000
```

Para los componentes principales, obtenemos pesos en el factor 1 practicamente iguales a los de los métodos **sin rotar** anteriores, de este modo, podemos mantener la interpretación ya realizada.
Para el factor 2, la variable con el peso de mayor magnitud es para el porcentaje de la población economicamente activa desocupada ($PeaDes$), las demás variables tienen pesos menos significativos, de este modo, el segundo factor caracteriza a la población desocupada.
Para el factor 3, obtenemos pesos similares obtenidos a los pesos en el factor 2 de los modelos sin rotación anteriores, entonces, la interpretación es similar, es decir, las variables de mayor peso son las dos últimas ($PeaSec$ y $PeaTer$) con signos contrarios, y esto puede significar un contraste entre ambos para zonas no rurales.
También notemos que las comunalidades son iguales a 1 en las cuatro variables, además, la proporción acumulada de la varianza es igual a 1, lo que es muy significativo a comparación de los métodos anteriores.

4.6 Método de Componentes Principales Con Rotación

Por último, el método de componentes principales pero con rotación varimax.

```
# Aplicamos el método de componentes principales y rotacion varimax con 3 factores
fit.m3.rot <- principal(R_lab, nfactors=3, rotate="varimax")
fit.m3.rot
```

```
 Principal Components Analysis
Call: principal(r = R_lab, nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
          RC1   RC3   RC2 h2      u2 com
PeaDes  -0.01  0.02  1.00  1 1.1e-16 1.0
PeaPrim -0.88 -0.48  0.00  1 1.5e-10 1.5
PeaSec   0.13  0.99  0.02  1 4.0e-11 1.0
PeaTer   1.00 -0.03 -0.02  1 1.0e-10 1.0

                       RC1  RC3  RC2
SS loadings           1.79 1.21 1.00
Proportion Var        0.45 0.30 0.25
Cumulative Var        0.45 0.75 1.00
Proportion Explained  0.45 0.30 0.25
Cumulative Proportion 0.45 0.75 1.00

Mean item complexity =  1.1
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0

Fit based upon off diagonal values = 1
```

En el factor 1, las variables con más pesos son $PeaPrim$ y $PeaTer$, lo cual obtuvimos en los modelos **con rotación** anteriores, entonces, podemos mantener la interpretación, es decir, el factor 1 contrasta la población ocupada en el sector primario (común en zonas rurales) y la población ocupada en el sector terciario (común en zonas no rurales), esto es parecido a lo que resultó en el modelo sin rotación.

El segundo factor (el cual es `RC2` pero está en la tercer columna), tenemos el mayor peso en la primer variable ($PeaDes$), lo cual coincide con lo que se obtuvo en el modelo sin rotar, de este modo, el factor 2 caracteriza a la población desocupada.

El tercer factor (el cual es `RC3` pero está en la segunda columna), se tiene el peso más grande en la variable $PeaSec$ (también es la variable con mayor peso del factor 3 en el modelo sin rotar), por tanto, el factor 3 se centra en caracterizar solo a la población no rural que se dedica al sector secundario.

5. Conclusiones

Haciendo una observación general a los distintos modelos ajustados, vemos que, para los tres subgrupos, el método de componentes principales ayudó a que se tuviera una buena comprensión de las cualidades que conforman un espacio rural y las que lo diferencian de un espacio no rural, ya que fue el modelo el cual tuvo interpretaciones similares a los otros métodos pero también con una mejor proporción acumulada de la varianza (en los tres caso mayores al 80%) y con comunalidades muy altas; sin embargo, también hay que dar crédito a los métodos de verosimilitud y de factores principales, ya que otorgaron una buena interpretación haciendo uso solamente de 2 factores e incluso 1 (en el caso del subgrupo Vivienda), mientras que el método de componentes usó 3 factores en todos los casos.
También, dando un resúmen de los análisis, se puede decir que en una zona o espacio rural tiende a caracterizarse por tener localidades chicas las cuales carecen de servicios como el drenaje y la electricidad, también es en estas zonas donde la población, por lo general, se dedica al sector primario (actividades agrícolas, pecuarias, silvícolas, etc.).

6. Referencias Bibliográficas

- Méndez, Roberto. (2009). ¿Qué es la ruralidad?. ACAbase: La Cooperación. Recuperado de http://portal.acabase.com.ar/lacooperacion/Lists/EntradasDeBlog/Post.aspx?ID=128.
- Prado, L. N. (1995). Introducción. Lo rural y la ruralidad: algunas reflexiones teóricometodológicas. El Colegio de Michoacán. Recuperado de https://www.colmich.edu.mx/relaciones25/files/revistas/054/LuzNereidaPerezPrado.pdf.