UNIVERSITAT JAUME I

Grado en Matemática Computacional

Estancia en Prácticas y Proyecto Final de Grado

Análisis estadístico de datos de planes de financiación universitarios

Autor:
Ana López Díaz

Supervisor:
Juan Antonio Hernández Rubert
Tutor académico:
María Victoria Ibáñez Gual
Amelia Simó Vidal

Fecha de lectura: __ de _____ de 2019
Curso académico 2018/2019

Resumen

Este documento corresponde al Trabajo de Fin de Grado presentado en la asignatura MT1030 - Prácticas Externas y Proyecto Final de Grado del grado en Matemática Computacional impartido por la Universitat Jaume I (UJI).

En el proyecto podemos encontrar el trabajo realizado en la estancia en prácticas en el Gabinete de Planificación y Prospectiva Tecnológica de la UJI. Este trabajo consiste en la aplicación del Plan Plurianual de Financiación del sistema de universidades públicas valencianas, así como en el análisis estadístico de las variables que intervienen en el plan.

Además, también se incluyen los fundamentos teóricos estadísticos que se han utilizado para desarrollar el trabajo. Se verá en detalle los métodos utilizados para la imputación de datos faltantes, el método de Análisis de Componentes Principales y los gráficos multivariantes propuestos.

Palabras clave

Plan de financiación, Imputación, Análisis de Componentes Principales, Gráficos Multivariantes.

Keywords

Financing plan, Data imputation, Principal component analysis, Multivariate plot

Índice general

Capítulo 1

Introducción

1.1. Contexto y motivación del proyecto

En el año 2010 se publicó el Plan Plurianual de Financiación para el período 2010-2017 del Sistema Universitario Público Valenciano (SUPV) que sustituiría el PPF anterior cuyo acuerdo se remonta a 1999. Los principales objetivos del nuevo plan de finaciación fueron establecer una estabilidad financiera, tanto para las Universidades Valencianas como para el Gobierno Valenciano, definir un criterio de asignación de recursos basado en diferentes ámbitos como son la docencia, investigación, etc. y por último, incitar a la reorientación y mejora de los servicios universitarios de acuerdo a la demanda.

Este modelo de financiación no se ha llegado a aplicar por parte del Gobierno Valenciano, y puesto que próximamente tendrá lugar una negociación para un nuevo plan de financiación, al Gabinete de Planificación y Prospectiva Tecnológica de la Universitat Jaume I le interesa conocer cómo queda situada la UJI al aplicar el PPF de 2010-2017 respecto al resto de universidades.

Es por eso que se han realizado simulaciones del presupuesto que obtendría cada universidad pública de la Comunidad Valenciana si se aplicara el PPF 2010-2017 y una comparativa gráfica de los resultados obtenidos. Se han estudiado también posibles modificaciones en el PPF actual con la introducción de nuevas variables y un análisis de las variables actuales para conocer cuales pueden favorecer más a la Universitat Jaume I y con qué universidades públicas de la Comunidad Valenciana tiene más parecido.

Capítulo 2

Estancia en prácticas

2.1. Introducción

En este capítulo se va a exponer el trabajo realizado durante la estancia en prácticas en el Gabinete de Planificación y Prospectiva Tecnológica de la Universidad Jaume I. Este organismo, situado en el edificio de Rectorado de la universidad, tiene como misión dar apoyo a los organismos de gobierno en la planificación universitaria, el desarrollo de proyectos institucionales, etc.

La estancia en prácticas en este gabinete consiste principalmente en el análisis, a través del uso de diferentes técnicas de minería de datos, del resultado de aplicar diferentes planes plurianuales de financiación a las universidades públicas de la comunidad valenciana, además de dar apoyo en la toma de decisiones relativa a la negociación de un plan nuevo de financiación.

2.2. Objetivos del proyecto formativo

Los objetivos que se pretenden conseguir con la realización de este proyecto son los siguientes:

- Hacer uso de los conocimientos adquiridos en la asignatura MT1045 - Sistemas de Apoyo a la decisión para poder analizar los indicadores utilizados en los diferentes PPF y realizar informes utilizando el programa Power BI.
- Uso de Excel para realizar las simulaciones de los distintos PPF.

- Adquisición de conocimientos en técnicas de minería de datos.
- Integración de la alumna en una primera experiencia laboral.
- Dar apoyo al Gabinete de Planificación y Prospectiva Tecnológica realizando simulaciones de la aplicación de distintos PPF a los datos de las diferentes universidades públicas de la Comunidad Valenciana para su posterior análisis y comparativa.

2.3. Metodología y definición de tareas

Las principales tareas que se han llevado a cabo se enumeran a continuación.

T1 Estudio del PPF del Sistema Universitario Público Valenciano (SUPV)

T2 Obtención de la información necesaria de las universidades del SUPV

T3 Simulación de la aplicación del PPF a las universidades del SUPV

T4 Búsqueda en internet de PPF de diferentes sistemas universitarios

T5 Simulación de la aplicación de los diferentes PPF a las universidades del SUPV

T6 Comparación de indicadores y resultados del PPF del SUPV y los PPF encontrados de otros sistemas universitarios

T7 Análisis de los resultados obtenidos y propuestas de mejora del PPF del SUPV.

2.3.1. Planificación temporal de las tareas

Durante la primera quincena de la estancia en el Gabinete de Planificación y Prospectiva Tecnológica de la UJI se ha trabajado sobre el Plan Plurianual de Financiación del Sistema Universitario Público Valenciano (de ahora en adelante, SUPV).
El trabajo realizado ha consistido principalmente en conocer los indicadores que se tienen en cuenta en el PPF mencionado anteriormente y obtener los datos necesarios para su aplicación. Se han utilizado además técnicas de imputación de datos para obtener valores faltantes. (Tareas 1 y 2)

Las tareas realizadas en la segunda quincena han consistido en la simulación del PPF del SUPV para todas las universidades públicas de la Comunidad Valenciana en el año 2018 y en específico para la UJI en el período de 2013-2019. Así como la generación de diferentes gráficos

para realizar una comparativa de las variables en la UJI y del presupuesto obtenido para cada universidad. (Tarea 3)

En el período de la tercera quincena, se han generado diferentes gráficos para realizar una comparativa de las variables y presupuesto obtenidos para la UJI durante el período 2013-2019 y también para visualizar las diferencias del presupuesto obtenido para cada universidad en un mismo año. (Tarea 3)

En la cuarta quincena se han realizado búsquedas tanto online como en diferentes libros con el fin de encontrar otros planes de financiación de universidades. A partir de los planes de financiación encontrados, se han realizado diferentes simulaciones donde se compara el uso del modelo del SUPV con el de otras universidades. (Tareas 4, 5 y 6)

En la quinta quincena se han estudiado los fundamentos teóricos del PCA y se ha aplicado a las variables del PPF con el fin de conocer relaciones entre universidades e intentar disminuir el número de variables. (Tarea 7)

Por último, en la sexta quincena se han realizado diferentes gráficos multivariantes a partir de todos los resultados obtenidos con el programa R. (Tarea 7)

2.4. Trabajo realizado

En esta parte encontraremos dos secciones. Una primera donde se explica a rasgos generales el funcionamiento del PPF del SUPV y se referencian los informes obtenidos a partir de su aplicación.

En la segunda parte encontraremos dos planes de financiación diferentes; el de las universidades de Andalucía y Madrid. De estos dos planes, se han tomado las características o variables más interesantes y se han aplicado al PPF del SUPV para una posterior comparativa.

2.4.1. PPF del SUPV

Descripción

Este apartado se corresponde con la tarea **T1** donde se procede a detallar el PPF del SUPV.

El modelo de financiación del Sistema Universitario Público Valenciano [1] toma en consideración los 3 bloques que se muestran en la figura 2.1.

Durante todo el proyecto nos vamos a referir a las diferentes partes que componen el plan de financiación según la siguiente notación:

- FR: Financiación por Resultados
- FE: Financiación Estructural
- FC: Financiación por mejora de la Calidad
- SRD: Subvención por Resultados Docentes
- SRI: Subvención por Resultados de Investigación
- SRT: Subvención por Resultados de Transferencia Tecnológica e Innovación
- SF: Subvención Fija
- SCI: Subvención por Capacidad Investigadora
- $SCNEA$: Subvención para la compensación de los Costes inducidos por la Normativa Estatal y Atonómica
- SMC: Subvención por la mejora de calidad

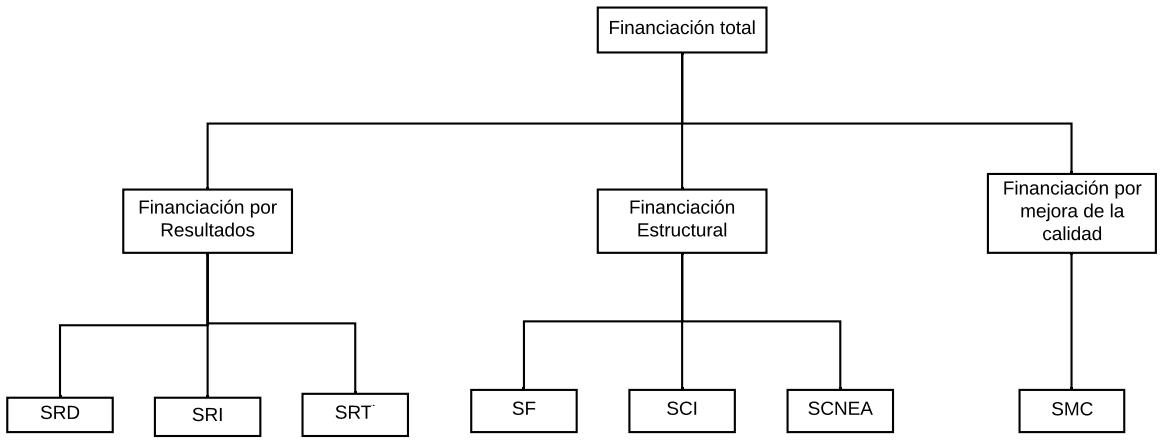


Figura 2.1: Estructura general del modelo de financiación para las universidades del SUPV

La financiación total anual FU para cierta universidad en un determinado año viene dada por la suma de los 3 bloques FR, FE y FC calculados para esta.

$$FU = FR + FE + FC$$

Procedemos a detallar el cálculo de cada uno de estos bloques.

Financiación por Resultados (FR)

La financiaicón por Resultados pretende cubrir las actividades básicas como son la docencia, investigación y transferencia de cada universidad en cada año. Esta subvención se divide en 3 bloques de acuerdo a la ecuación indicada a continuación.

$$FR = SRD + SRI + SRT$$

1. **Subvención por Resultados Docentes (SRD)**

   Por una parte tenemos la Subvención por Resultados Docentes (SRD), basada en los costes de la docencia por estudiante matriculado a tiempo completo (ETC), definido como aquel matriculado de 60 créditos. Entonces, la SRD de una universidad a percibir en cierto año, se puede calcular como:

   $$SRD = \sum_j snetc_j \cdot NETC_j$$

   donde j hace referencia a cada titulación
   Donde $NETC$ se refiere al número de estudiantes a tiempo completo y $snetc$ es la subvención neta por ETC de cada titulación,

   $$NETC = \frac{CrdMat}{60}$$

   $$snetc = 0{,}73 \cdot cmetc$$

   Para calcular la subvención neta por estudiante a tiempo completo para cada titulación, necesitamos calcular primero el coste estándar de los servicios docentes por ETC ($cmetc$), formado por el coste medio del personal docente e investigador ($cmpdi$), el coste medio del personal de administración y servicios ($cmpas$), el gasto corriente en bienes y servicios

($cmgg$), el gasto corriente en prácticas docentes ($cmpd$) y las inversiones en infraestructuras y equipamientos ($cminv$) según la ecuación:

$$cmetc = \beta(cmpdi + cmpas + cmgg + cminv) + cmpd$$

β es un coeficiente que se corresponde con los gastos docentes sobre los totales y se estima como 0.7279. Su complementario se corresponde con la actividad investigadora.

Es importante tener en cuenta que tanto el $cmppdi$, $cmppas$, $cmgg$, $cmpd$ y $minv$ se actualizarán cada año a partir del PIB nominal, siguiendo la fórmula

$$coste\ actual = coste\ 2010 \cdot (1 + PIBNominal)$$

Coste medio del personal docente e investigador ($cmpdi$)

Para calcular el coste medio del personal docente e investigador se multiplica el número de profesores necesarios por ETC ($NPDI$) y el coste medio de la plantilla de profesorado ($cmppdi$). El coste medio de la plantilla de profesorado se estima para 2010 en 46.062,62 €. El número de profesores necesarios por ETC se calculará como

$$NPDI = \frac{60 \times exp}{22 \times TMGT}$$

siendo exp el coeficiente de experimentalidad asociado a esa titulación y $TMGT$ el tamaño medio de los grupos de teoría.

$$cmpdi = cmppdi \cdot NPDI$$

Coste medio del personal de administración y servicios ($cmpas$)

El coste medio del personal de administración y servicios se obtiene multiplicando el número de PAS por estudiante a tiempo completo ($NPAS$) y el coste medio de la plantilla de PAS ($cmppas$). Para calcular el $NPAS$ se utiliza la relación $\frac{PDI}{PAS} = 2,2$ corregida con el nivel de experimentalidad de la titulación para la cual estamos realizando el cálculo, quedando

$$NPAS = NPDI \cdot \frac{(1 + \frac{exp}{6})}{2,2}$$

Así, el coste medio del personal de administración y servicios sigue la fórmula

$$cmpas = cmppas \cdot NPAS$$

Coste medio asociado al gasto corriente en bienes y servicios ($cmgg$)

Se obtiene a partir de la siguiente ecuación

$$cmgg = mgca \cdot NPDI$$

donde $mgca$ hace referencia al gasto corriente por alumno matriculado a tiempo completo, el cual se calcula para 2010 en 3.345.43 €.

Coste medio asociado al gasto corriente en prácticas docentes ($cmpd$)

Es calculado mediante

$$cmpd = mgcp \cdot exp$$

siendo $mgcp$ el gasto corriente por profesor, estimado como 446.06 € en 2010, y exp el coeficiente de experimentalidad correspondiente a la titulación sobre la que se realiza el cálculo.

Coste de las inversiones en infraestructuras y equipamientos ($cminv$)

El valor correspondiente a las inversiones en infraestructuras y equipamientos vendrá dado según la experimentalidad de la titulación.

$$cminv = minv \cdot exp$$

El valor de las inversiones que deben hacerse para mantenimiento de construcciones, instalaciones y equipos ($minv$) para 2010 se ha estimado como 223.03 €.

2. **Subvención por resultados de Investigación (SRI)**

La subvención por resultados de Investigación para una universidad viene dada según la siguiente expresión:

$$SRI = (snac \times NAC) + (snsex \times NSEX) + (snric \times NRIC)$$

Las variables implicadas en este cálculo son el número de artículos científicos en la Web of Science de los últimos tres ejercicios (NAC), el número de sexenios aprobados en los últimos seis ejercicios ($NSEX$) y la media de recursos públicos captados en investigación de los últimos tres ejercicios ($NRIC$).
Siendo *snac*, *snex* y *snric* las subvenciones unitarias para cada una de estas variables, cuyo valor se puede consultar en la tabla 2.1.
Estos valores se actualizarán anualmente según el PIB y un coeficiente α que viene definido por el coste que representa la parte de la jornada que puede dedicar a la investigación la plantilla de PDI con capacidad investigadora. Sin embargo, en las simulaciones que se han realizado en los últimos años y durante este proyecto se ha mantenido el valor de α como 1, por lo que los valores de las subvenciones unitarias se actualizan unicamente según el PIB.

3. **Subvención por resultados de transferencia tecnológica e innovación (SRT)**

Para calcular la subvención por resultados de transferencia tecnológica e innovación, utilizaremos la expresión mostrada a continuación:

$$SRT = (snrctr \times NRCTR) + (snfc \times NRFC)$$

Las variables $NRCTR$ y $NRFC$ hacen referencia a la media de ingresos captados a partir de contratos con empresas en los últimos tres años y a la media de ingresos por cursos de formación continuada en los últimos tres ejercicios, respectivamente. Las subvenciones unitarias que corresponden a estas dos variables; $snrctr$ y $snfc$ se pueden consultar en la tabla 2.1.

Tipo	Subvención unitaria 2010 en €
snac	1.284,74
snsex	6.025,70
snric	0,28
snrctr	0,38
snfc	0,34

Tabla 2.1: Subvenciones unitarias.

Financiación Estructural (FE)

La Financiación Estructural FE viene determinada por la siguiente expresión

$$FE = SF + SCI + CNEA$$

1. **Subvención Fija**

   La financiación estructural consta de una primera subvención fija, cuyos valores se pueden consultar en la tabla 2.2

Universidad	Subvención fija 2010 en €
UA	3.953.851,91
UJI	5.930.777,87
UMH	5.930.777,87
UPV	3.953.851,91
Uv	3.953.851,91

Tabla 2.2: Subvencion fija 2010.

   Para actualizar esta parte de la subvención se hará teniendo en cuenta la previsión de variación del PIB nominal, según la fórmula

$$subvención\ fija = subvención\ 2010 \times (1 + previsión\ variación)$$

2. **Subvención por capacidad investigadora**

   La ecuación utilizada para el cálculo de la subvención por capacidad investigadora

$$SCI = \alpha \times cmcipdi \times NETC$$

   tiene en consideración el factor α comentado anteriormente, cuyo valor estará entre 0 y 1 con la finalidad de darle más o menos importancia a los resultados en I+D+i a la hora de financiar la actividad investigadora. En principio se considera un factor α=1, el cual se irá reduciendo con el tiempo aumentando así las subvenciones unitarias, tratadas en el apartado anterior, siguiendo la expresión

$$\Delta(snac, snsex, snric, snrcrt, snfc) = 1{,}0934\Delta\alpha$$

Para realizar el cálculo de las subvenciones unitarias por resultados al modificar el factor α se seguirá la ecuación

$$subvención\ actual = subvención\ 2010 \times (1 + variación\ deflactor\ PIBNominal) \times (1 + (1 - \alpha) \cdot 1{,}0934)$$

Aunque, como ya habíamos comentado, se ha mantenido el valor de α como 1 durante todo el proyecto.

Otro elemento que se tiene en cuenta a la hora de calcular la subvención por capacidad investigadora es el coste de la capacidad investigadora de la plantilla de PDI por estudiante ETC (*cmcipdi*), cuyo valor depende del grado de experimentalidad y tipo de estudios, y se puede consultar en las tablas 2.3 y 2.4

Nivel de Experimentalidad	cmcipdi en €
Nivel 1	543,80
Nivel 2	582,65
Nivel 3	673,28
Nivel 4	747,96
Nivel 5	800,17

Tabla 2.3: cmcipdi para grado.

Nivel de Experimentalidad	cmcipdi en €
Nivel 1	1.196,37
Nivel 2	1.281,82
Nivel 3	1.631,41
Nivel 4	1.673,36
Nivel 5	1.999,64

Tabla 2.4: cmcipdi para máster.

3. **Subvención para la Compensación de los costes inducidos por la Normativa Estatal y Autonómica (CNEA)**

Esta subvención está destinada a cubrir aquellos gastos sobre los que las universidades no tienen capacidad de decisión, como son:

a) Coste por Complementos de Méritos Docentes del profesorado

b) Coste de las retribuciones por antigüedad de la totalidad del personal

c) Coste por Complementos de Productividad investigadora

d) Coste S. Social por transformación de contratos administrativos en laborales por adaptación LOU.

e) Coste por Complementos Cuerpos Docentes no Universitarios RD 74/2000

f) Coste por Retribuciones adicionales autonómicas PDI Decreto GV 174/2002

g) Coste por incremento del Complemento de Destino en pagas extraordinarias

h) Coste por incremento del Complemento Específico en pagas extraordinarias

i) Exenciones familias numerosas

j) Exenciones matrículas de honor

k) Exenciones discapacidad

l) Exenciones víctimas bandas armadas

m) Pago a Tribunal PAU mayores 25 años

n) Pago a Tribunal PAU LOGSE

$\tilde{n}$) Pago a Profesores Especialistas LOGSE

o) Pago a Profesores Asesores LOGSE

Financiación para mejorar la Calidad (FC)

El último bloque de subvenciones surge con la intención de hacer mejoras en la calidad de la actividad universitaria. Para ello se determinan unos objetivos que deberán ser cumplidos por la universidad para obtener la subvención ideada para ellos.
Como no se han llegado a desarrollar estos objetivos, durante todas las simulaciones se va a utilizar un valor fijo, el calculado para el año 2010 y que se puede consultar en la tabla 2.5.

Universidad	Subvención Objetivos 2010 en €
UA	7.321.381
UJI	3.715.236
UMH	4.250.666
UPV	14.312.394
UV	14.643.293

Tabla 2.5: Subvención Objetivos 2010.

Simulaciones

Una vez conocido el funcionamiento del plan de financiación del sistema universitario público valenciano, se han realizado diferentes simulaciones (tarea **T3**) en Excel tanto para la Universitat Jaume I como para el resto de universidades de la comunidad Valenciana. Es decir, se ha aplicado el modelo propuesto en el PPF para conocer el presupuesto que habría correspondido a cada universidad.

Para ello primero se ha realizado una búsqueda de los datos necesarios, que se corresponde con la tarea **T2**, para llevar a cabo estas simulaciones los cuales están disponibles en las páginas web del SIUPV [2], IVIE [3] y en el informe 'La universidad española en cifras' proporcionado por la CRUE [4].

A continuación en la tabla 2.6 se muestra un análisis descriptivo de los datos de las variables con los que vamos a trabajar durante el proyecto.

Variable	Media	Desv. Típica	Máx.	Mín.
NRFC	2.976.759,82	3.178.264	8.411.539,1	39.343
NRCTR	11.593.483,73	14.125.460,7	54.717.434,6	1.813.340,5
NRIC	21.372.929,52	16.572.350	53.652.406	3.794.906,5
GRADO1	143.088,38	145.606,7	433.333,5	66
GRADO2	390.316,13	309.038,9	926.902,2	79.182
GRADO3	117.164,27	74.825,38	241.116,5	6.681,3
GRADO4	334.965,35	279.401,3	1.011.301,8	122.331
GRADO5	126.863,77	139.571,6	873.659	0
MASTER1	22.209,65	17.025,73	61.020	7.301
MASTER2	41.567,63	31.396,31	106.522,5	10.567,5
MASTER3	13.307,9	10.243,26	33.916,5	2.471
MASTER4	34.896,78	34.272,24	123.943,5	6.939,7
MASTER5	7.164,78	6.604,479	19.960,5	0
SEX	151,6	84,569	277	52
NAC	3.373,76	2.291,252	8.322	986

Tabla 2.6: Análisis descriptivo de las variables.

La primera simulación ha consistido en aplicar el PPF del SUPV para los datos del período de 2013 a 2018 de la Universitat Jaume I. Se ha realizado así la comparación entre la subvención real proporcionada a la universidad y la calculada según el modelo. Los resultados se encuentran en las figuras A.1, A.2 y A.3 del Anexo I.

En la figura A.1 nos encontramos con varios gráficos. Un primer gráfico dónde se muestra

la evolución del número de estudiantes a tiempo completo (ETC) matriculados en la UJI para el periodo 2012-2019. Desde el año 2014 se observa un decrecimiento del valor de esta variable. Como consecuencia, la subvención calculada por el PPF para el bloque de Resultados docentes ha disminuido también en este período más de 10.000 €.

En cuanto a las variables que intervienen en el bloque de financiación, vemos que tanto las publicaciones referenciadas en WoS como los sexenios han experimentado un crecimiento durante todo el intervalo de tiempo estudiado, excepto en los sexenios del último año donde se puede ver una ligera disminución. En cuanto a los recursos captados, pese a que sufrieron una caída desde el año 2012 hasta el 2015, han presentado gran crecimiento hasta 2019 superando casi en 5.000 € el valor presentado por esta variable en 2012. Como en dos de tres variables ha incrementado el valor de la variable con el paso de los años, también lo ha hecho la subvención calculada, presentando un mínimo en el año 2015 que coincide con el año en que menos ingresos por investigación se captaron.

En la figura A.2 continua el informe correspondiente a esta simulación. La gráfica correspondiente a la subvención por Transferencia Tecnológica sigue la misma forma que la gráfica con los datos del presupuesto captado en contratos. La media de recursos captados en formación sigue la misma tendencia aunque más suave.

En la Financiación Estructural vemos que viene principalmente marcado por la CNEA, ya que en cuanto a la capacidad investigadora no se perciben grandes cambios.

Finalmente, en la figura A.3 vemos que en la Financiación por Resultados y Estructural, toma más importancia; casi un 70 %, la financiación por resultados frente a la estructural. Así mismo, la docencia toma gran protagonismo en la Financiación por resultados, seguido de la investigación y muy por detrás la transferencia tecnológica. En la Financiación Estructural parece estar más dividido entre la CNEA y la capacidad investigadora.

En la última gráfica se presenta el valor de la subvención calculada por el PPF y el valor de la subvención real recibida por la universidad en el paso de los años. Mientras que el valor calculado siguiendo el plan ha ido aumentando con los años, la subvención recibida ha sido prácticamente constante, presentando un máximo en el año 2016 y creciendo ligeramente a partir del año 2017. Aún y así, las diferencias entre ambos cálculos llega a los 20 millones de €.

Posteriormente, se ha realizado la simulación del PPF del SUPV al resto de universidades de la Comunidad Valenciana para el año 2017, con el fin de comparar los valores que toman los indicadores usados así como la financiación que corresponde a cada una de las universidades. Los informes realizados se pueden consultar en la figura A.4 del Anexo I.

Al comparar a la UJI con el resto de universidades del SUPV, vemos que tanto la UJI como la UMH son las únicas que reciben menos presupuesto del que correspondería según el PPF del

SUPV.

2.4.2. Otros PPF

Después de la aplicación del PPF del SUPV, se ha pasado a la búsqueda de otros PPF de diferentes universidades (tarea **T4**). Es importante mencionar que los PPF encontrados de otras universidades son planes competitivos, es decir, en lugar de basar su modelo en calcular la cantidad de dinero a repartir, lo que se hace es calcular porcentajes a partir de ciertas variables y/o indicadores para obtener el % de la financiación que se dispone que le corresponde a cada universidad.

Por los problemas surgidos a la hora de obtener los datos necesarios para la aplicación de los PPF que mencionaremos a continuación, se ha optado por aplicar simplemente aquellas variables que se consideren más interesante y/o novedosas y analizar las diferencias al incluirlas en el PPF del SUPV.

PPF de Andalucía

Descripción

El plan de financiación de la comunidad de Andalucía tiene la estructura general que se muestra en la figura 2.2.

La **Financiación Ordinaria Básica** se refiere a aquella destinada a cubrir el desarrollo operativo ordinario de las universidades. Esta se puede separar en 3 bloques: Estructura institucional, Oferta académica y Demanda docente.

1. La **Financiación básica de carácter estructural** es la que pretende cubrir los gastos de la universidad que radican en su propia existencia. Esta financiación será equitativa para todas las universidades de la comunidad de Andalucía.
2. La **Financiación básica vinculada a la oferta académica** intenta cubrir los costes de implantación y mantenimiento que suponen las diferentes titulaciones ofertadas. Esta parte de la financiación se distribuirá de acuerdo con el perfil de cada universidad, el cual depende de la duración de obtención del título y del nivel de experimentalidad. En el cálculo de esta parte del presupuesto sólo intervien dos variables: Los coeficientes asociados a la duración y experimentalidad y el número de titulaciones de cada nivel de experimentalidad. El producto del coeficiente y el número de titulaciones de ese nivel nos determinará un valor para cada universidad, que será equivalente a su oferta. Realizando

una agregación de los valores obtenidos para todas las universidades podremos conocer el valor global de la oferta total, el cual corresponderá al 7,5 % de la financiación total.

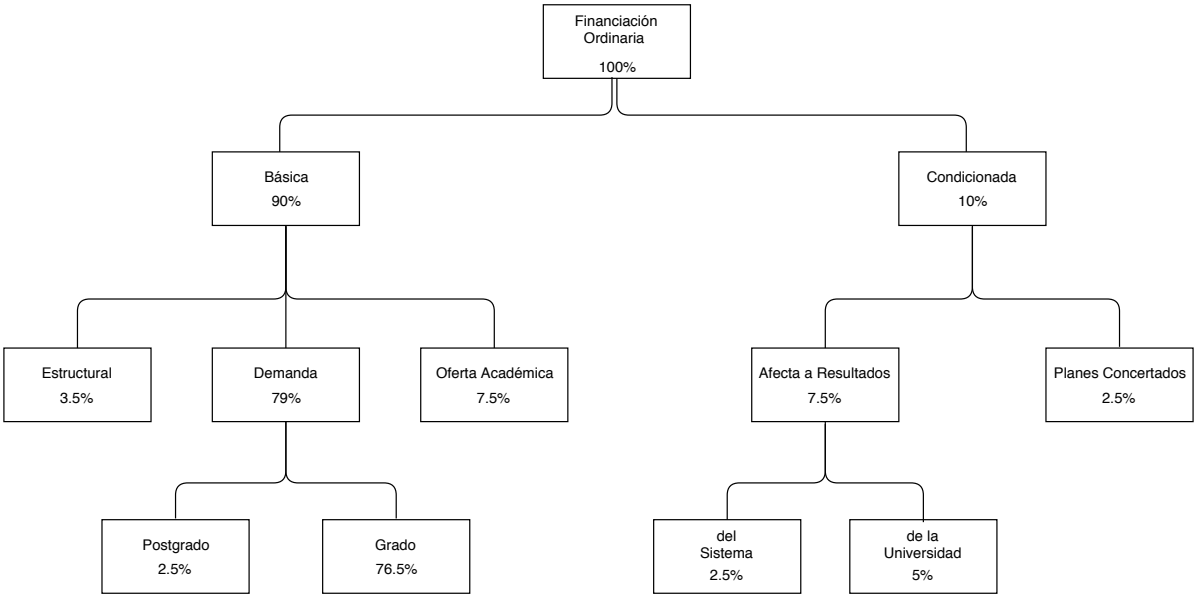


Figura 2.2: Estructura general del modelo de financiación para las universidades de Andalucía

3. En la **Financiación básica vinculada a la demanda de servicios docentes universitarios** a diferencia de en el PPF del SUPV, aquí se intenta normalizar el número de ETC (Estudiantes a tiempo completo) según la opción de matrícula (1ª, 2ª...) con la siguiente ecuación:

$$N_j = N_{j1} + 0{,}85N_{j2} + 0{,}8N_{j3}$$

siendo,
N_j = Alumnos de la asignatura j normalizados
N_{j1} = Alumnos de la asignatura j en opción de 1ª matrícula
N_{j2} = Alumnos de la asignatura j en opción de 2ª matrícula
N_{j3} = Alumnos de la asignatura j en opción de 3ª matrícula

La **Financiación ordinaria condicionada** se estructura en dos bloques:

Un primer bloque, la **Financiación para Planes Concertados**, que se calcula a partir de pogramas de prácticas de empresas, convocatorias de ayudas, etc.
Y un segundo bloque, la **Financiación afecta a Resultados** que viene derminada por la consecuación de determinados objetivos y logro de ciertos resultados que acuerden las Universidades.

Simulación

Debido a la falta de información sobre el cálculo de los diferentes presupuestos y la dificultad a la hora de obtener los datos necesarios, la única simulación que se ha podido realizar ha sido incorporar la idea del PPF de Andalucía de valorar la oferta académica al PPF del SUPV.

Para ello, vamos a mantener las experimentalidades establecidas para las universidades de la Comunidad Valenciana. Para realizar el cálculo de la oferta académica de cada universidad, se realiza un recuento de todas las titulaciones que comparten el mismo nivel de experimentalidad y se aplica el coeficiente definido para tal nivel. El valor de la oferta académica para una universidad se correspondería a la agregación de los resultados obtenidos. Estos valores se han calculado mediante una tabla dinámica y podemos observarlos en la tabla 2.7.

En esta misma tabla, además del valor de la oferta de cada universidad aparece la diferencia de presupuesto al incluir esta subvención en el PPF del SUPV del ejercicio de 2018.
Para poder realizar esta comparación se ha partido de la idea del PPF de Andalucía de que la subvención por la oferta académica sea un 7.5 % del total y se ha seguido el siguiente procedimiento:
Por una parte tenemos el valor de la oferta académica de cada universidad y su valor total. Con lo cual podemos calcular el % del valor de la oferta que tiene cada universidad.

Por otra parte disponemos de la subvención para cada universidad y también la subvención total destinada a todas las universidades, 88.750.576,98 €, ya que es la suma de cada una de estas.
Como el 7,5 % del presupuesto va destinado para la oferta, calculamos el 7,5 % del valor del presupuesto total destinado a las universidades. De ahí obtenemos el valor del presupuesto destinado para la oferta académica, 44.156.293,27 €.

También, disponiendo del valor de la subvención para cada universidad, podemos calcular el % que le corresponde a cada universidad. Es decir, como el 7,5 % del presupuesto total destinado para las universidades equivale a 44.156.293,27 €, al restarle este valor a la subvención total, obtenemos la parte de la subvención que nos queda por repartir, 544.594.283,71 €, y que repartiremos en base al % de la subvención que tenía cada una de las universidades.

Así, hemos redistribuido el presupuesto destinado a las universidades teniendo en cuenta la oferta académica.

A pesar de ser la UPV y UV las universidades con mayor oferta académica, son las que salen perdiendo en comparación al PPF original, mientras que universidades pequeñas como la

UJI y UMH verían incrementada su subvención.

Universidad	Valor Oferta	Diferencia respecto al presupuesto original (en €)
UA	238,84	1.238.084,94
UJI	135,93	365.165,93
UMH	156,51	1.045.944
UPV	249,57	-923.828,16
UV	345,33	-1.725.366,71

Tabla 2.7: Valor oferta académica universidades SUPV

PPF de Madrid

Descripción

El plan de financiación de las universidades de la Comunidad de Madrid toma en consideración 3 bloques de financiación que podemos ver la figura 2.3.

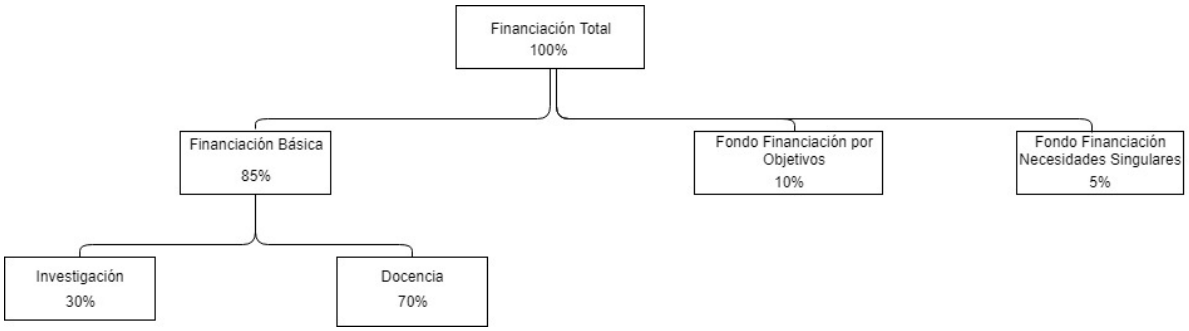


Figura 2.3: Estructura general del modelo de financiación para las universidades de Madrid

El **Fondo de Financiación Básica** tiene el objetivo de financiar las actividades relacionadas con la docencia y la investigación.

Para realizar la distribución de la **Financiación Básica para la Docencia** se toma en cuenta la valoración de la función docente. Esto se calcula teniendo en cuenta 2 variables, la capacidad instalada (tomada como el coste medio total estándar destinado a docencia) y el número de estudiantes a tiempo completo (ETC). El producto de estos dos valores teniendo en cuenta la experimentalidad (ver tabla 2.9) nos da como resultado la valoración de la función docente. La agregación de la valoración de todas las universidades corresponderá al 70 % de la financiación total.

El reparto de la **Financiación Básica para la Investigación** se realiza de forma análoga al de docencia, pero tomando esta vez como variables el coste medio total etándar destinado a investigación y el número de invesigadores equivalentes de la universidad (NIE), teniendo también en cuenta la experimentalidad del área al que pertencen.
El cálculo del NIE se hace mediante la valoración de la actividad de la plantilla del PDI la cual se calcula mediante el producto del % en obtenido en cada indicador y su ponderación. Los indicadores son los siguientes:

P1 = (Sexenios reconocidos/Sexenios posibles) x 100
P2 = (Fondos externos investigación/Ingresos corrientes) x 100
P3 = (Tesis leídas/PDI doctores a tiempo completo) x 100
P4 = (Becas/Plantilla PDI) x 100
P5 = (Fondos externos art. 83 LOU/Ingresos corrientes) x 100
P6 = Publicaciones científicas (Queda por determinar su cálculo)

Indicador	Coeficiente
P1	0,5
P2	0,25
P3	0,05
P4	0,05
P5	0,15
P6	Sin determinar

Tabla 2.8: Coeficientes de los indicadores de investigación Madrid

El siguiente bloque que nos encontramos es el **Fondo de Financiación por Objetivos**, distrubución la cual se hace de acuerdo al cumplimiento de objetivos de cada universidad. Agrupando los objetivos en 8 áreas, tenemos los siguientes indicadores para calcular la valoración de este cumplimiento.

Objetivo 1: Área Estratégica de Reestructuración de la Oferta Docente

I1 = (Preinscritos 1ª opción/Oferta) x 100
I2 = (Matrícula N.I 1ª opción/Matrícula N.I) x 100
I3 = (Matrícula N.I de fuera de la comunidad/Matrícula N.I) x100
I4 = (Matrícula N.I/Oferta) x100
I5 = (Matrícula de N.I/Admitidos) x100
I6 = (Matrícula N.I de extranjeros/Matrícula N.I) x100
I7 = (Ingresos matrícula reglada/Ingresos totales) x100

Objetivo 2: Área Estratégica de Mejora en el Rendimiento de las Actividades Docentes

I8 = % de alumnos que terminan estudios en el tiempo teórico de duración
I9 = % de mejora en la reducción del número medio de años de permanencia
I10 = (Créditos aprobados/Créditos matriculados) x100 I11 = (Titulados/Matrícula N.I) x100

Objetivo 3: Área Estratégica de Inserción Laboral de los Titulados

I12 = % de titulados con empleo en el 3er año posterior a la titulación

Objetivo 4: Área Estratégica de Renovación Pedagógica y Nuevas Tecnologías

I13 = (Créditos prácticos en empresas/Créditos totales impartidos) x100
I14 = % de inversiones en tecnologías de la información respecto de los gastos totales

Objetivo 5: Área Estratégica de Cualificación de Plantillas y Cobertura de Créditos Matriculados

I15 = (Nº de PDI tiempo completo/Nº total de PDI) x100
I16 = (Nº de PDI doctores/Nº total de PDI) x100
I17 = (Total de créditos matriculados/Matrícula total) x100 x 1.54

Objetivo 6: Área Estratégica de Actividades de Formación Permanente

I18 = % de ingresos por matrícula no reglada respecto de los ingresos totales por matrícula

Objetivo 7: Área Estratégica de Mejora de la Calidad de los Servicios

I19 = % valoración externa de la calidad de la universidad

Objetivo 8: Área Estratégica de Resultados de la Investigación

I20 = (Sexenios reconocidos al PDI/ Sexenios posibles del PDI) x100
I21 = (Fondos externos art.83 LOU/Ingresos totales) x100
I22 = (Tesis leídas/PDI doctores a tiempo completo) x100

El último bloque, el **Fondo de Financiación de Necesidades Singulares** está destinado a cubrir los gastos corrientes y compromisos ineludibles que no se corresponen con ninguno de los bloques anteriormente descritos.

Experimentalidad	Coeficiente
1	1,56
2	1,51
3	1,47
4	1,3
5	1,17
5bis	1,09
6	1

Tabla 2.9: Tabla de experimentalidades Madrid.

Simulación

La primera simulación basada en este PPF ha sido aplicar las experimentalidades propuestas, que se pueden ver en la tabla 2.9, al cálculo de la SRD del PPF del SUPV.
Los coeficientes de experimentalidad del SUPV varían desde 1,40-2,34 para máster y 1,40-2,06 para grado. Al ser mayores que los coeficientes de Madrid (van desde 1,00 hasta 1,56), el SRD total resultante será menor que si tomáramos las experimentalidades del SUPV.

Aún así, si realizamos comparaciones porcentuales sobre la cantidad que le correspondería a cada universidad, no se observan grandes variaciones. Ver las figuras A.6 y A.7

La segunda simulación realizada consiste en el cálculo de los indicadores de la Financiación Básica para la Investigación. En este caso se han realizado 3 pruebas diferentes.
Una primera donde se ha tomado el porcentaje del PPF de Madrid que irá destinado a investigación (30 % del 85 % del total, es decir, un 25,5 % del total) y se ha aplicado al presupuesto total que se ha invertido en el SUPV en el año 2017.

Una seguna prueba donde se ha tomado la cantidad total aportada a investigación por la Generalitat en el año 2017 calculada a partir de la suma del presupuesto en investigación, es decir, $SRI + SRT$. Esa cantidad se reparte entre las diferentes universidades del SUPV.

La tercera prueba toma el presupuesto total indicado en la segunda prueba y le añade también el valor de la subvención SCI.

Para repartir el presupuesto entre las universidades, se ha incorporado parte del modelo de financiación básica para la investigación de la Comunidad de Madrid.

Estos indicadores son ligeramente diferentes a los propuestos por Madrid.
Por una parte, se ha eliminado el indicador P4, ya que no se disponía de los datos necesarios para calcularlo.

Por otra parte, el indicador P6 del PPF de Madrid (Publicaciones científicas) no estaba definido, así que se ha tomado la relación entre el número de publicaciones del último año y los PDI a tiempo completo. Como coeficiente para este indicador se tomará el valor del indicador eliminado P4.
Además, los indicadores P2 y P5 se calculaban respecto a los ingresos corrientes. En cuanto a ese dato, sólo disponemos del calculado en el PPF del SUPV, siendo el mismo para todas las universidades. Si usásemos ese valor para calcular el coeficiente no saldría un reparto justo, así que se ha decidido tomar en su lugar los ingresos totales de cada universidad.

Luego los indicadores quedarían como se muestra a continuación:

P1 = (Sexenios reconocidos/Sexenios posibles) x 100
P2 = (Fondos externos investigación/Ingresos totales) x 100
P3 = (Tesis leídas/PDI doctores a tiempo completo) x 100
P4 = (Fondos externos art. 83 LOU/Ingresos corrientes) x 100
P5 = (Publicaciones científicas/PDI a tiempo completo) x 100

A cada indicador se le asigna una ponderación, mostrada en la tabla 2.8, y luego se traslada la puntuacuón de cada universidad a una escala de forma que la suma de sus porentajes nos de como resultado el 100 %.

Así, los indicadores y sus respectivos coeficientes quedarían como se muestra en la tabla

Indicador	Coeficiente
P1	0,5
P2	0,25
P3	0,05
P4	0,05
P5	0,05

Tabla 2.10: Modificación coeficientes de los indicadores de Investigación Madrid

Una vez calculados los porcentajes, realizamos el reparto del presupuesto en investigación.

Los resultados a rasgos generales indican que la subvención sería mayor que la original para las universidades más pequeñas, como la UJI o la UA, mientras que las más grandes sufrirían una pérdida respecto al original. En el anexo se pueden consultar los 3 informes; A.5, A.8 y A.9, realizados a partir de los resultados obtenidos.

En la figura A.5 podemos ver los valores que toman los indicadores propuestos por el PPF de Madrid en las diferentes universidades del SUPV. En este caso la UJI presenta un resultado muy bueno en el indicador relacionado con los sexenios, colocándose en segunda posición precedido de la UV. En cuanto a fondos de investigación y tesis leídas se encuentra la última en el ranking, aunque en esta última no presenta valores tan alejados del resto de universidades. En el resto de indicadores; fondos externos y publicaciones científicas se encuentra en un tercer puesto. La mayor diferencia la encontramos en fondos externos donde la UPV abarca más del 50 % del indicador.

Por otra parte, en la figura A.8 podemos visualizar una comparativa del valor de la subvención recibida por cada universidad del SUPV según las 3 diferentes aplicaciones realizadas y comentadas anteriormente. En todas ellas la UJI obtiene valores casi similares a la UMH, siendo estas dos las que menos presupuesto recibirían, pero aún así, mayor que con el PPF original.

Por último, en la figura A.9 se presenta la diferencia de subvención entre aplicar las 3 simulaciones planteadas con los indicadores del PPF de Madrid y el PPF del SUPV. En todas ellas son las universidades más pequelas (UJJI y UMH) las favorecidas económicamente, mientras que las universidades más grandes salen más perjudicadas, destacando la UV.

Optimización

Posteriormente, se han planteado estas operaciones como ecuaciones de un problema de optimización y se han calculado los coeficientes que harían que la UJI obtuviera el mayor presupuesto posible.
Siendo c_1, c_2, c_3, c_4 y c_5 los coeficientes a encontrar y 0,21, 0,138, 0,172, 0,13 y 0,2 el valor obtenido para cada indicador respecto al total de universidades. 150131397,13 es el presupuesto total a repartir. La función a maximizar sería:

$$f(c_1, c_2, c_3, c_4, c_5) = 150131397,13 \cdot (0{,}21c_1 + 0{,}138c_2 + 0{,}172c_3 + 0{,}13c_4 + 0{,}2c_5)$$

Sujeto a

$$c_1 + c_2 + c_3 + c_4 + c_5 = 1$$

$$0{,}05 \leq c_1, c_2, c_3, c_4, c_5 \leq 0{,}8$$

Para ello, en primer lugar, se ha dado el mismo valor a todos los indicadores: 0,2.
Usando la herramienta SOLVE de Excel se ha intentado maximizar el valor del presupuesto destinado a la UJI, tomando como valor inicial 0,2 para todos los coeficientes.
Los resultados obtenidos han sido los mostrados en la tabla 2.11. La conclusión obtenida para esta primera prueba es que el indicador que toma más importancia para la UJI es el P1.

Indicador	Coeficiente
P1	0,8
P2	0,05
P3	0,05
P4	0,05
P5	0,05

Tabla 2.11: Coeficientes de los indicadores de investigación Madrid que maximizan el problema

En segundo lugar se propone la misma función a maximizar con unas condiciones un poco más estrictas:

$$máx\, f(c_1, c_2, c_3, c_4, c_5) = 150131397,13 \cdot (0{,}21c_1 + 0{,}138c_2 + 0{,}172c_3 + 0{,}13c_4 + 0{,}2c_5)$$

Sujeto a
$c_1 + c_2 + c_3 + c_4 + c_5 = 1$
$0{,}05 \leq c_1, c_2, c_3, c_4, c_5 \leq 0{,}4$

Así, el resultado obtenido ha sido el mostrado en la tabla 2.12. De aquí podemos asumir, que el segundo indicador que más puede favorecer a la UJI es el P5, el cual se basa en el número de artículos publicados. El tercero que más favorece es P3, basado en las tesis leídas. Seguido de P2 y P4, ambos basados en ingresos.

Indicador	Coeficiente
P1	0,4
P2	0,05
P3	0,1
P4	0,05
P5	0,04

Tabla 2.12: Coeficientes de los indicadores de investigación Madrid que maximizan el problema

Finalmente, se han aumentado las restricciones haciendo que los coeficientes varíen entre

0.05 y 0.25. De aquí obtenemos los coeficientes que harían que la UJI obtuviera el máximo beneficio en el bloque de investigación usando los indicadores mencionados.

Indicador	Coeficiente
P1	0,25
P2	0,2
P3	0,25
P4	0,05
P5	0,25

Tabla 2.13: Coeficientes indicadores investigación Madrid que maximizan el problema

2.5. Grado de consecución de los objetivos propuestos

Por lo general, los objetivos del proyecto se han podido satisfacer en su mayoría. Si bien no ha sido posible realizar un análisis para poder resumir y disminuir el número de variables que intervienen el el PPF del SUPV, sí que ha sido posible hacer una comparativa de todas las universidades aplicando el plan actual. También se ha podido analizar las variables usadas y la incorporación de otras nuevas, así como la posición de la UJI respecto al resto de universidades.

2.6. Conclusiones

La experiencia de la estancia en prácticas en el Gabinete de Planificación y Prospectiva Tecnológica me ha acercado a una primera experiencia laboral muy satisfactoria. Con la ayuda de mi supervisor he podido aprender todo lo necesario que requería la puesta en marcha del proyecto.

Por otra parte, he adquirido conocimientos sobre técnicas de análisis, imputación de datos y gráficos multivariantes gracias a mis tutoras que me han proporcionado toda la información necesaria para realizar este proyecto.

Capítulo 3

Fundamentación Teórica del TFG

3.1. Motivación y Objetivos de la Minería de datos

La minería de datos [5] nace con el objetivo de ayudar a comprender grandes cantidades de datos y así extraer conclusiones de ellos, contribuyendo en el crecimiento y la mejora de las empresas. La finalidad principal es la exploración de bases de datos grandes automáticamente, utilizando diferentes tecnologías y técnicas estadísticas, hasta encontrar tendencias, reglas o patrones en los datos.

Dada una muestra de n elementos en los que hemos medido p características o variables $(X_1, X_2, ..., X_p)$,podemos representar estos valores en forma de una matriz de datos $\mathcal{X}^{n\times p}$.

$$\mathcal{X}^{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

La cual también se puede expresar utilizando el vector observación de cada individuo.

$$\mathcal{X}^{n\times p} = \begin{pmatrix} \mathbf{x_1}^t \\ \mathbf{x_2}^t \\ \vdots \\ \mathbf{x_n}^t \end{pmatrix}$$

Siendo

$$\mathbf{x_i} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

el vector observación en el individuo i-ésimo.

3.2. Imputación de Datos

En las bases de datos es frecuente encontrarse con la existencia de datos faltantes o *missing values*. Esto se puede deber a múltiples motivos, entre ellos errores en la transcripción de los datos o directamente la ausencia de respuesta.

Antes de realizar un análisis sobre un muestreo de datos o aplicar diferentes técnicas estadísticas es necesario tratar la ausencia de datos, ya que estos pueden derivar en una pérdida de validez y/o disminución de la información de la muestra.

Cuando tenemos una muestra de datos univariante, la imputación se puede realizar únicamente teniendo en cuenta los datos observados de esa variable. El método más conocido para estos casos es la imputación de la media.
Por otra parte, cuando manejamos una muestra multivariante nos podemos servir de las observaciones del resto de variables para imputar nuestros datos faltantes.

En este proyecto, como ya se ha visto, tenemos una muestra multivariante y un único valor faltante: La variable NRCTR para la Universitat Jaume I en el año 2012. Nos vamos a servir de este ejemplo para ver diferentes técnicas de imputación restringiéndonos a aquellas implementadas por el programa R [6].
Para ello, procederemos a quitar un dato conocido y obtener el error del dato imputado. Construiremos un bucle que realice esto con cada dato de la muestra y calcularemos la media del error y la media del error al cuadrado, para así decidir que método de imputación utilizar.

3.2.1. imputePCA

Esta función pertenece al paquete de R *missMDA* [7]. Su función es imputar los valores faltantes de la muestra utilizando el algoritmo iterativo de PCA. Las bases teóricas del PCA se

explicarán en el siguiente apartado. Para realizar la imputación se siguen los siguientes pasos [8]:

1. Los datos faltantes de la matriz de datos $\mathcal{X}$ se rellenan con la media de la variable correspondiente.
2. Para cada iteración h se aplica PCA en la matriz completa para estimar los parámetros M^h, F^h y U^h. M^h es una matriz de dimensión $n \times p$ donde cada fila contiene la media de cada variable, U^h es la matriz de coeficientes, es decir, la matriz de vectores propios. Y $\mathcal{Z}^h$ es la matriz de puntuaciones, de forma que la varianza de cada columna es igual al valor propio correspondiente.
3. Se construye la matriz $\hat{\mathcal{X}}^h = M^h + {U^h}^t \cdot \mathcal{Z}^h$ a partir de 3.5. Y siendo W una matriz con valor 0 en aquellas posiciones que $\mathcal{X}$ tenga datos faltantes y valor 1 en las posiciones con valores observados, la nueva matriz imputada quedará definida por
$$\mathcal{X}^h = W \times \mathcal{X} + (1 - W)\hat{\mathcal{X}}^h$$
4. Se repite los pasos 2-3 hasta conseguir la convergencia de las variables imputadas.

3.2.2. missForest

missForest [9] está basado en bosques aleatorios, una técnica de aprendizaje supervisado donde se combinan árboles de decisión sobre el mismo conjuntos de datos.

Un árbol de decisión [10] es un grafo $G = (V, A)$ donde dos vértices cualquiera están unidos por un único camino. Además, asumimos que es un grafo binario dirigido (es decir, cada nodo tiene dos nodos hijos) con un nodo raíz.
El árbol de decisión tiene como función predecir el valor de una variable ayudándose del resto de variables contenidas en nuestra muestra.

Podemos encontrarnos con dos tipos de árboles de decisión: Aquellos donde la variable a predecir (o variable respuesta) puede tomar un conjunto finito de valores, conocidos como árboles de clasificación, y aquellos donde la variable respuesta puede tomar valores continuos, que serán los utilizados para los datos de este proyecto, los árboles de regresión.

Para construir un árbol de regresión [11] se comienza seleccionando una variable X_i sobre la cual obtenemos un punto de corte c_i, de manera que separe por un lado los datos con $X_i > c_i$ y por otra aquellos con $X_i < c_i$. Esta condición definirá el nodo inicial. En cada uno de los dos nuevos nodos que parten del inicial se aplicará el mismo proceso hasta tener todos los datos clasificados. En la figura 3.1 podemos ver un ejemplo de árbol de decisión. La forma de seleccionar las variables y los cortes se basa en criterios frecuencialistas y se pueden consultar en [11].

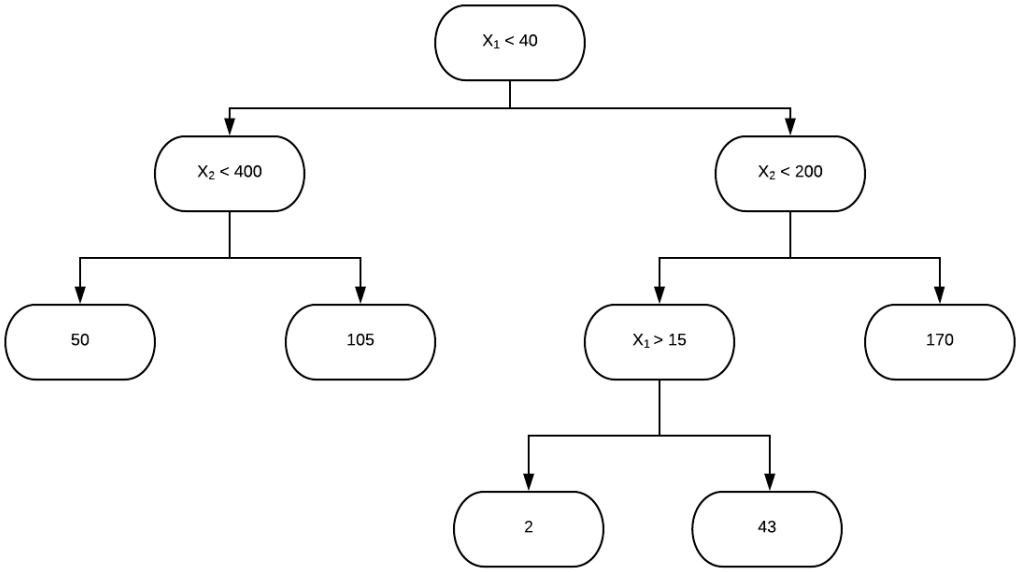


Figura 3.1: Ejemplo árbol de decisión

Por otra parte, un bosque aleatorio o *random forest* es un conjunto de árboles de decisión creados a partir de un subconjunto aleatorio (con repetición) de la muestra del problema con la intención de combinar sus resultados utilizando, por ejemplo, la media.

Las bases teóricas de los árboles de decisión y los bosques aleatorios quedan fuera del alcance de esta memoria, pero pueden consultarse en [12].

Los pasos a seguir del algoritmo [13] son los siguientes:

1. Estimación inicial para los datos faltantes utilizando la media de cada variable o cualquier otro método de imputación.
2. Se ordenan las variables X_s; $s = 1, 2, ...p$, según la cantidad de datos faltantes en cada una de ellas, de menor a mayor. Siendo p el número total de variables.
3. Para cada variable X_s se crea un bosque aleatorio con los datos conocidos de la muestra y se predice la parte faltante.
4. La ejecución se repite hasta llegar a la condición de parada. El criterio de parada se define de forma que la ejecución finaliza cuando la diferencia entre la matriz nueva y la imputada previamente aumenta por primera vez.

La diferencia para el conjunto de las variables $X_1, X_2, ...X_i$ se define como

$$\Delta X = \frac{\sum_{i,j=1}^{i=n,j=p} (\hat{x}_{ij}^{nuevo} - \hat{x}_{ij}^{viejo})^2}{\sum_{i,j=1}^{i=n,j=p} (\hat{x}_{ijnueva})^2}$$

Aplicación

Para ver como funcionan estos dos métodos con nuestros datos programamos un bucle en el que, en cada iteración eliminamos un dato (no faltante) de una variable de nuestro fichero de datos, lo etiquetamos como desconocido y lo imputamos para obtener el error entre el valor original y el valor calculado con cada uno de los métodos.

Al ejecutar el bucle utilizando la función *imputePCA* para calcular la media de error de cada dato, el resultado ha sido

$$\frac{1}{n \cdot p} \sum_{i,j=0}^{i=n,j=p} \varepsilon_{ij} = 1739004$$

y la media del error al cuadrado

$$\frac{1}{n \cdot p} \sum_{i,j=0}^{i=n,j=p} \varepsilon_{ij}^2 = 1{,}058496 \cdot 10^{15}$$

siendo $\varepsilon_{ij} = |x_{ij} - \hat{x}_{ij}|$ el error calculado para la imputación del dato x_{ij}.

Y al ejecutarlo con la función *missForest* la media de error de cada dato ha sido

$$\frac{\sum_{i,j=0}^{i=n,j=p} \varepsilon_{ij}}{n \cdot p} = 2110739$$

y la media del error al cuadrado

$$\frac{\sum_{i,j=0}^{i=n,j=p} \varepsilon_{ij}^2}{n \cdot p} = 1,6255 \cdot 10^{15}$$

3.3. Análisis de Componentes Principales

3.3.1. Teoría

Frecuentemente los datos que queremos analizar se caracterizan por contener un gran número de variables. Pese a que estas variables nos pueden aportar información valiosa, un gran número de ellas puede dificultar su análisis. Es por eso que reducir la dimensión (el número de variables) sin que se produzca una gran pérdida de información puede resultar muy útil para el análisis estadístico de datos.

PCA (Análisis de Componentes Principales) es una técnica estadística que permite resumir la información contenida en una muestra de datos con muchas variables. Su objetivo es representar los datos originales como un conjunto de nuevas variables incorreladas entre sí, llamadas componentes principales.

Siguiendo la notación anterior, sea $\mathcal{X}$ nuestra matriz de datos originales, con n observaciones y p variables

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

$$\mathcal{X}^{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\mathcal{X}^{n\times p} = \begin{pmatrix} \mathbf{x_1}^t \\ \mathbf{x_2}^t \\ \vdots \\ \mathbf{x_n}^t \end{pmatrix}$$

S será su matriz de varianzas y covarianzas cuadrada y simétrica de orden p. Esta matriz contiene en su diagonal las varianzas de las variables y en el resto de posiciones las covarianzas entre ellas. Se define como:

$$S = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^t \tag{3.1}$$

siendo $\bar{\mathbf{x}}$ el vector de medias de cada variable.

En cambio, si la matriz de datos está centrada (es decir, si su vector de medias es 0) podremos escribir S a partir de $\mathcal{X}$ como

$$S = \frac{1}{n}\mathcal{X}^t\mathcal{X} \tag{3.2}$$

Lo que busacmos son $k < p$ variables nuevas $(Z_1, Z_2, ..., Z_p)$, combinación lineal de las originales y incorreladas entre sí a las cuales llamaremos componentes principales.

Al ser combinación lineal de las variables originales podemos definir las nuevas variables como

$$Z_i = \mathbf{u_i}^t X$$

y el vector observación i-ésimo:

$$\mathbf{z_i} = \mathcal{X} \cdot \mathbf{u_i} \tag{3.3}$$

Queremos perder la menor cantidad de información posible, por lo que buscamos que la varianza de las nuevas variables Z_i sea máxima. La varianza muestral de la variable Z_i, S_{Z_i} se puede escribir como

$$S_{Z_i} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{z_{ji}}^2 = \frac{1}{n}\mathbf{z_i}^t \cdot \mathbf{z_i}$$

que aplicando la igualdad 3.3 nos quedará

$$S_{Z_i} = \frac{1}{2}\mathbf{u_i}^t \cdot \mathcal{X}^t \cdot \mathcal{X} \cdot \mathbf{u_i}$$

donde podemos aplicar la definición de la matriz S dada en 3.2

$$S_{Z_i} = \mathbf{u_i}^t \cdot S \cdot \mathbf{u_i} \tag{3.4}$$

Para encontrar la primera componente principal Z_1 buscamos maximizar 3.4 con la restricción de que los vectores sean unitarios. Es decir, se debe cumplir

$$\mathbf{u_i}^t \mathbf{u_i} = 1$$

Aplicamos el método de los multiplicadores de Lagrange a la función

$$L = \mathbf{u_1}^t S \mathbf{u_1} - \lambda(\mathbf{u_1}^t \mathbf{u_1} - 1)$$

para afrontar tanto la maximización como la restricción. Derivando respecto al vector $\mathbf{u_1}$ e igualando a 0 obtendremos

$$\frac{\delta L}{\delta \mathbf{u_1}} = 2S\mathbf{u_1} - 2\lambda \mathbf{u_1} = 0$$

$$S\mathbf{u_1} = \lambda \mathbf{u_1}$$

Lo que significa que λ es un valor propio de la matriz S. Si ahora multiplicamos a ambos lados por $\mathbf{u_1}^t$

$$\mathbf{u_1}^t S \mathbf{u_1} = \mathbf{u_1}^t \lambda \mathbf{u_1}$$

aplicando la ecuación 3.4 llegamos a

$$S_{Z_i} = \lambda_1$$

Esto quiere decir, que para maximizar la varianza de $\mathbf{z_1}$ habrá que tomar el mayor valor propio de la matriz S y el vector propio asociado a este como $\mathbf{u_1}$.

Es decir, la primera componente principal vendrá dada por

$$\mathbf{z_1} = \mathcal{X} \cdot \mathbf{u_1}$$

Para construir el resto de componentes principales se añade la restricción de que deben estar incorreladas entre sí, es decir, deben formar una base ortonormal.

La segunda componente principal vendrá dada por el segundo mayor autovalor asociado a la matriz S y su vector propio correspondiente $\mathbf{u_2}$

$$\mathbf{z_2} = \mathcal{X} \cdot \mathbf{u_2}$$

Y así análogamente hasta obtener todas las componentes principales.

$$\mathcal{Z} = \mathcal{X}U \tag{3.5}$$

Siendo U la matriz de vectores propios o direcciones principales.

Pero, ¿Qué pasa cuando $n < p$? Este es el caso con el que nos hemos encontrado en este proyecto. Cuando $n < p$, la matriz S será singular. Esto supone un problema ya que aparecen valores propios nulos y puede aparecer un problema computacional a la hora de calcular el valor de aquellos no nulos.

La solución a esto es utilizar la Descomposición en Valores Singulares (SDV).

Los valores singulares de una matriz $\mathcal{X}$ se definen como la raíz cuadrada de los valores propios de $\mathcal{X}\mathcal{X}^t$

Aplicando la Descomposición en Valores Singulares de la matrix de datos $\mathcal{X}$ ya centralizada (es decir, habiendo restando las medias a cada columna), podemos descomponerla en otras tres matrices.

$$\mathcal{X} = UDV^t$$

Siendo r el rango de la matriz $\mathcal{X}$, $U_{n\times r}$ es la matriz que contiene los vectores propios asociados a los valores propios de $\mathcal{X}\mathcal{X}^t$, $V_{r\times p}$ la matriz que contiene los vectores propios asociados a $\mathcal{X}^t\mathcal{X}$ y $D_{r\times r}$ la matriz diagonal con los valores singulares.

Luego la matriz de varianzas covarianzas S (3.2) se puede escribir como

$$S = \frac{1}{n}VDU^tUDV^t = V\frac{D^2}{n}V^t$$

lo que significa que los vectores singulares V de la matriz X son los vectores propios de nS y que los valores singulares γ_i están relaciones con los valores propios de la matriz de varianzas-covarianzas mediante la siguiente expresión

$$\lambda_i = \frac{\gamma_i^2}{n}$$

3.3.2. Aplicación

Uno de los objetivos iniciales de la estancia en prácticas fue estudiar si se podía disminuir el número de variables del PPF, por ello en este apartado procedemos a realizar dos análisis de componentes principales que nos servirá para ello y para poder conocer la situación de la UJI respecto al resto de universidades. Para el primero de ellos vamos a tomar las variables que intervienen en el PPF del SUPV previamente estandarizadas en el intervalo [0,1].

Las variables utilizadas son las siguientes:

- X_1=NRFC: Media de euros captados en formación continua en los últimos tres años.
- X_2=NRCTR: Media de euros captados en contratos con empresas en los últimos tres años.
- X_3=NRIC: Media de euros captados en programas de investigación en los últimos tres años.
- X_4=GRADO1: Media de número de créditos matriculados en grados de experimentalidad 1 en los últimos cuatro años.
- X_5=GRADO2: Media de número de créditos matriculados en grados de experimentalidad 2 en los últimos cuatro años.
- X_6=GRADO3: Media de número de créditos matriculados en grados de experimentalidad 3 en los últimos cuatro años.
- X_7=GRADO4: Media de número de créditos matriculados en grados de experimentalidad 4 en los últimos cuatro años.
- X_8=GRADO5: Media de número de créditos matriculados en grados de experimentalidad 5 en los últimos cuatro años.
- X_9=MASTER1: Media de número de créditos matriculados en máster de experimentalidad 1 en los últimos cuatro años.
- X_{10}=MASTER2: Media de número de créditos matriculados en máster de experimentalidad 2 en los últimos cuatro años.
- X_{11}=MASTER3: Media de número de créditos matriculados en máster de experimentalidad 3 en los últimos cuatro años.
- X_{12}=MASTER4: Media de número de créditos matriculados en máster de experimentalidad 4 en los últimos cuatro años.
- X_{13}=MASTER5: Media de número de créditos matriculados en máster de experimentalidad 5 en los últimos cuatro años.
- X_{14}=NAC: Total de artículos referenciados WoS en los últimos tres años.
- X_{15}=SEX: Total de sexenios reconocidos en los últimos tres años.

Con el comando **pca <- prcomp(variables)** se realiza el análisis de componentes principales sobre los datos de la tabla *variables*. Este comando nos devuelve un objeto de la clase *prcomp* con varias componentes. Una de ellas contiene el valor de los loadings para cada componenete. Es decir, los vectores propios. Podemos visualizarla utilizando el comando **pca$rotation**.

	PC1	PC2	PC3	PC4	PC5
NRFC	0.279381064	-0.33101913	-0.16728198	0.51504159	0.15072482
NRCTR	0.033036566	-0.42527910	-0.05647986	-0.09305714	-0.49729054
NRIC	0.229771224	-0.33438714	0.03059889	-0.13156525	0.24183259
GRADO1	0.312755678	0.19183909	0.41226399	-0.15260065	0.43346809
GRADO2	0.270671086	0.16878273	0.51146387	0.06223323	0.01230952
GRADO3	0.291432213	-0.02608575	-0.08058635	-0.20028456	-0.27510581
GRADO4	0.009440038	-0.37725804	0.05658175	-0.25627217	0.21680329
GRADO5	0.336702790	0.19290958	-0.16948874	-0.16668305	-0.10132216
MASTER1	0.292475417	0.13111110	-0.07731894	0.38244065	-0.04736437
MASTER2	0.293844186	0.16422774	0.06103444	-0.17458008	-0.21593893
MASTER3	0.321119431	-0.08971015	-0.16782740	0.32842446	0.04290786
MASTER4	0.053641113	-0.38771077	-0.11766716	-0.28731682	0.38276325
MASTER5	0.239812694	0.31334149	-0.62519371	-0.32664980	0.16201598
SEX	0.264184565	-0.18953310	0.23404039	-0.25206739	-0.35490834
NAC	0.305623096	-0.12492840	0.02354561	0.10314997	-0.03272739

Figura 3.2: Resultado del comando pca$rotation

Por otra parte, utilizando el comando **pca$x** obtenemos el valor de las componenentes principales para cada observación.

	PC1	PC2	PC3	PC4	PC5
UA	-0.4573347	0.3501671	0.578120780	-0.18646247	-1.101549e-16
UJI	-0.7445253	0.3676757	0.081096713	0.30250432	4.579670e-16
UMH	-0.5110749	0.5951637	-0.593419262	-0.13796860	-2.081668e-16
UPV	-0.1092992	-1.6208191	-0.073201609	-0.01526165	3.122502e-17
UV	1.8222341	0.3078126	0.007403378	0.03718840	8.673617e-17

Figura 3.3: Resultado del comando pca$x

Realizamos el cálculo de varianza para cada componente y lo representamos gráficamente obteniendo la gráfica A.10

Como podemos observar en el gráfico A.10, con las dos primeras componentes tendríamos aproximadamente el 90 % de la varianza explicada.

Si ahora volvemos a mirar los resultados obtenidos en la figura 3.2, podemos seleccionar aquellas variables que mejor resuman los dos primeros componenetes principales.
En cuanto al primer componente podríamos decir que aquellas variables que mejor lo resumen

(ordenadas de mayor a menor) son **GRADO5, MASTER3, GRADO1, NAC, MASTER2, NRFC, SEX...** aunque la mayoría tienen valores bastante parecidos. Podemos decir que las variables **GRADO4, MASTER4 y NRCTR** apenas tienen peso en esta componente.

Siguiendo la misma línea, en la segunda componente principal tampoco encontramos unas pocas variables que destaquen por encima del resto. El valor más alto que encontramos es el perteneciente a **NRCTR**, a diferencia de la primera componenete en la cual tenía muy poco peso. **GRADO4 y MASTER4** son las siguientes variables con más peso.

Si ahora analizamos los datos de la figura 3.3 vemos por ejemplo, que la UPV tiene un valor alto en PC2, lo cual significa que sus valores de **GRADO4 y MASTER4** son elevados. Esto se debe a que la oferta de esta universidad es en mayor parte ingenierías y otros grados de experimentalidad 4. Por otra parte, vemos que la UV tiene el valor más bajo para PC2, lo cual vendría a explicar lo contrario.

De esta forma podemos ver en qué variables destacan más las universidades. Ahora vamos a proceder a poner en forma de gráfico los datos anteriores en la figura A.11. En esta figura, encontramos en el eje de las X las puntuaciones de la primera componente principal, mientras que en el eje de las Y se reflejan las puntuaciones de la segunda componente principal. Se han representado las diferentes universidades y las variables en función del peso que tienen en cada componente.

Ahora vamos a realizar las misma pruebas pero convirtiendo las variables del PPF del SUPV en indicadores relativos. A continuación se explica cómo se ha construido cada indicador:

- NRFCi = NRFC/ingresos totales
- NRCTRi = NRCTR/ingresos totales
- NRICi = NRIC/ingresos totales
- GRADO1i = GRADO1/total créditos de grado matriculados
- GRADO2i = GRADO2/total créditos de grado matriculados
- GRADO3i = GRADO3/total créditos de grado matriculados
- GRADO4i = GRADO4/total créditos de grado matriculados
- GRADO5i = GRADO5/total créditos de grado matriculados
- MASTER1i = MASTER1/total créditos de máster matriculados
- MASTER2i = MASTER2/total créditos de máster matriculados

- MASTER3i = MASTER3/total créditos de máster matriculados
- MASTER4i = MASTER4/total créditos de máster matriculados
- MASTER5i = MASTER5/total créditos de máster matriculados
- SEXi = SEX/sexenios reconocidos
- NACi = NAC/PDI doctores a tiempo completo

Realizando los mismos pasos que para las variables anteriores, obtenemos los resultados que se muestran e las figuras 3.4, 3.5, A.12 y A.13.

	PC1	PC2	PC3	PC4	PC5
NRFC	0.34052939	0.043017493	0.354683138	0.27659149	0.03612372
NRCTR	0.35394007	0.031813040	-0.004964784	-0.04769754	0.09422045
NRIC	0.21051519	0.019010119	0.047930930	0.35963691	0.28873932
GRADO1	-0.37722800	0.247550640	0.031029899	0.52493100	-0.19495250
GRADO2	-0.29087507	0.382085389	0.170125750	-0.16235787	-0.22993670
GRADO3	0.02704906	-0.420968531	-0.018488242	0.11925795	-0.12444508
GRADO4	0.34308332	0.007164121	-0.114094325	-0.15511056	-0.30283839
GRADO5	-0.20763393	-0.568588114	0.005340786	0.30459016	-0.02012293
MASTER1	-0.21802718	0.023846584	0.429513496	-0.41377542	0.51041426
MASTER2	-0.29797651	0.094353087	-0.200420403	0.20660415	0.18399868
MASTER3	0.11309019	-0.053720215	0.675266572	0.05869734	-0.45277119
MASTER4	0.34408468	0.029535271	-0.168069648	0.07662722	0.19395101
MASTER5	-0.22015229	-0.487346599	0.010157786	-0.28787387	-0.09199265
SEX	-0.09751392	0.088898116	0.167591493	0.17744401	0.21627026
NAC	0.02247388	-0.174092689	0.302509088	0.14609456	0.34019814

Figura 3.4: Resultado del comando pca$rotation

	PC1	PC2	PC3	PC4	PC5
UA	-0.5385956	0.581550444	-0.57069338	0.04658745	-2.775558e-17
UJI	-0.3820486	0.329225887	0.38453380	-0.34269398	1.009433e-15
UMH	-0.2989612	-0.999425278	-0.20431386	-0.08175369	4.510281e-16
UPV	1.7148278	0.083306869	-0.01274769	0.02572325	-2.602085e-17
UV	-0.4952224	0.005342078	0.40322113	0.35213697	-1.158795e-15

Figura 3.5: Resultado del comando pca$x

Observando la figura A.13 podemos ver que, en cuanto al componente principal PC1 todas las universidades son bastante parecidas excepto la UPV que toma un valor mucho más elevado. En cuanto a la segunda componente principal las universidades UV y UPV son las más similares

entre sí. Luego tenemos a la UJI con un valor ligeramente más alto que las dos anteriores y la UA que toma el mayor valor para esta componente. La UMH sin embargo, se encuentra alejada de todas ellas con el menor valor.

Fijándonos en las variables que aparecen en esta figura o bien, mirando los datos de la figura 3.2 podemos ver qué variables influyen más en cada componente y por tanto, en qué variables encontramos más similitudes o diferencias en cuanto a las universidades.

Analizando primero la primera componente principal, donde todas las variables tenían un valor negativo excepto la UPV, podemos ver que se podría resumir en las variables **GRADO4, MASTER4**, y mayoritariamente también en los ingresos. Lo cual nos dice que la UPV destaca en estudios de experimentalidad 4 y en ingresos. Por otra parte aquellas universidades con valores negativos para PC1 podríamos decir que destacan en estudios de experimentalidad 1 y 2.

Pasando ahora a la segunda componente principal, UV y UPV eran las universidades más parecidas con valores positivos pero próximos al 0. La UMH se encontraba alejada con un valor negativo para esta componente, lo cual representa que esta universidad ha conseguido un buen indicador para los estudios de experimentalidad 5.

3.4. Gráficos Multivariantes

Cuando nos encontramos con un conjunto de datos bivariante, podemos visualizar facilmente la relación entre las dos variables con simples gráficos 2D, como por ejemplo un diagrama de dispersión. Cuando en lugar de dos variables nos encontramos con tres, podemos visualizar la información también con un diagrama de dispersión o cualquier otro tipo de gráfico en 3D.

Sin embargo, plasmar en un plano la relación entre más de 3 variables; es decir, tres dimensiones, ya no es algo tan intuitivo. Aquí es dónde surgen los gráficos multivariantes [14], los cuales nos ayudan a estudiar y analizar conjuntos de datos cuando nos encontramos con un mayor número de variables.

En este proyecto se han utilizado tres tipos diferentes de representación de datos multivariante para conseguir una representación gráfica del conjunto de datos compuesto por las variables que intervienen en el PPF del SUPV.

3.4.1. Faces

La función faces [15] del paquete de R aplpack [16] nos muestra los datos a través de rostros. Cada variable representa una característica del rostro.

En nuestro caso, las variables que definen cada parte del rostro serían las siguientes.

- Altura de la cara: NRFC
- Ancho de la cara: NRCTR
- Forma del rostro: NRIC
- Altura de la boca: GRADO1
- Ancho de la boca: GRADO2
- Sonrisa: GRADO3
- Altura de los ojos: GRADO4
- Ancho de los ojos: GRADO5
- Altura del peinado: MASTER1
- Ancho del peinado: MASTER2
- Estilo de peinado: MASTER3
- Altura de la nariz: MASTER4
- Ancho de la nariz: MASTER5
- Ancho de las orejas: SEXENIOS
- Altura de las orejas: NAC

Al introducir el comando

```
faces(variables)
```

obtenemos el resultado mostrado en la figura 3.6.

Los colores utilizados en el rostro también nos indican relaciones entre las observaciones y las variables. Para escoger el color de cada característica del rostro se realiza la media de un

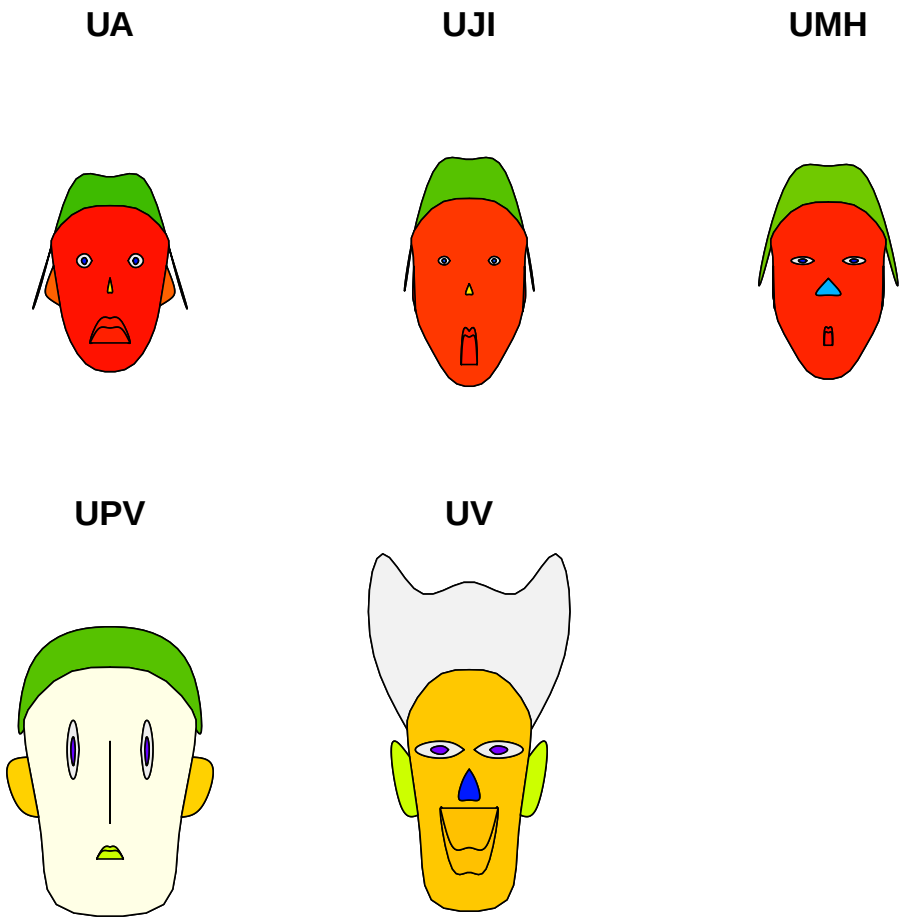


Figura 3.6: Resultado de aplicar la función *faces*

conjunto de variables. En este caso, el color de los ojos viene dado por el resultado de la media de las variables GRADO4 y GRADO5, que son las que nos definen el tamaño de los ojos. Para el color de los labios se realiza la media de las variables NRFC, NRCTR y NRIC. Por otra parte, el color de las orejas viene dado por la variable SEXENIOS y NAC. La nariz se colorea en función de la media de MASTER4 y MASTER5. El color del pelo se calcula en base al valor de las variables MATER1, MASTER2 y MASTER3. Y por último, el tono del rostro se consigue calculando la media de las variables NRFC y NRCTR.

La primera impresión es que las universidades UA, UJI y UMH son las más parecidas entre ellas, ya que tienen varias características en común: El ancho del rostro, la altura del rostro, el estilo del peinado, el tamaño de los ojos y el color del rostro, labios y peinado. Por otra parte, pese a que las universidades UPV y UV tienen el tamaño del rostro más grande que el resto de universidades (esto se debe a que tienen mayor cantidad de ingresos), entre ellas tienen características completamente opuestas.

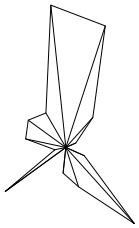
Podríamos decir que las tres universidades con un número parecido de ingresos son la UA, UJI y UMH. De entre ellas la UJI destacaría por créditos matriculados de experimentalidad 1 (Altura de la boca) y la UA sin embargo por créditos matriculados de experimentalidad 2 (Ancho dela boca). En cuanto a estas dos variables la UMH se queda por atrás, pero podemos ver que en cambio, supera en créditos matriculados de máster 4 y 5 a las otras dos universiades. En cuanto a GRADO4 y GRADO5 (ancho y altura de los ojos), la UA parece estar más proporcionada, la UJI mantiene la proporción pero con menor valor y la UMH tiene más valor en los créditos matriculados de GRADO5.

3.4.2. Diagrama estrella

Un diagrama en estrella [17] es otro método gráfico de mostrar datos en un espacio bidimensional con tres o más variables cuantitativas. Cada variable se representa como un segmento de tamaño proporcional a la magnitud de la muestra que corresponde. Es decir, cada segmento del diagrama representa una columna de la matriz de datos y el tamaño de la variable (escalada) se corresponde con el radio del segmento que representa a esa variable.

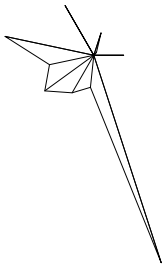
Utilizando la función *stars* [18] del paquete de R *graphics* [19], obtenemos el resultado de la figura 3.7

```
stars(variables)
```

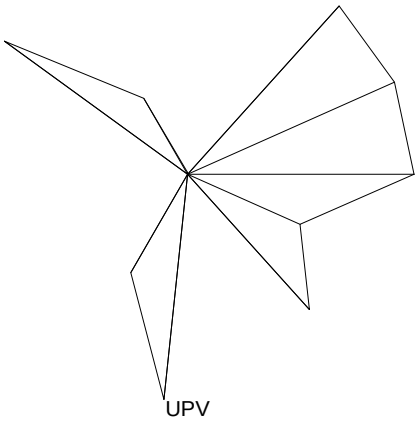


UA

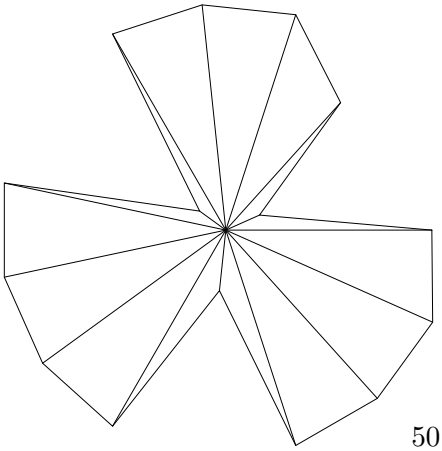
UJI



UMH



UPV



UV

Figura 3.7: Resultado de aplicar la función "stars".

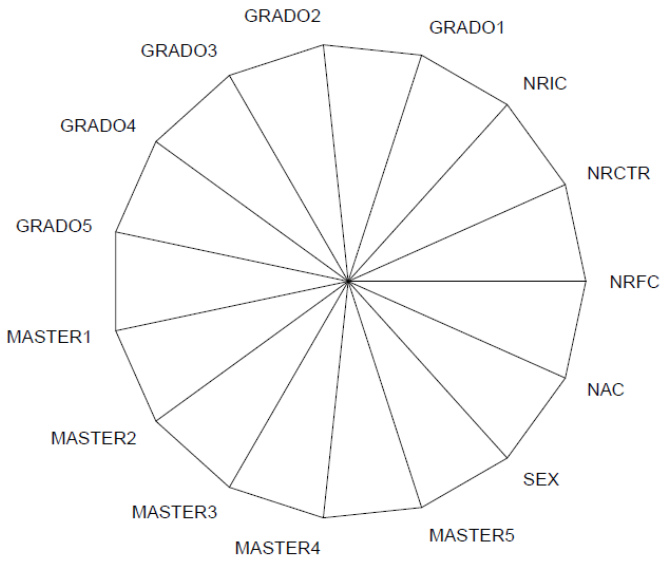


Figura 3.8: Variable que corresponde a cada segmento.

En la figura 3.8 podemos ver a qué variable corresponde cada segmento. Las variables están estandarizadas en el intervalo [0,1] por lo tanto aquellas observaciones donde la longitud del segmento sea mayor significará que esa variable toma su máximo valor en esa observación.

Por ejemplo, tanto la UA como la UJI no tienen el valor máximo para ninguna de las variables. Sin embargo, la UMH tiene el mayor valor posible para la variable MASTER5, la UPV tiene los máximos valores de las variables GRADO4, NRIC, NRCTR, NRFC y MASTER4. Y por último la UV es la universidad que más valores máximos toman sus variables; es decir, es la que tiene los valores más altos para el mayor número de variables: GRADO3, GRADO2, GRADO1, NAC, MASTER5, SEXENIOS, MASTER3, MASTER2, MASTER1 y GRADO5. También podemos ver que así como la UJI guarda relación con la UA en cuando a GRADO1 y GRADO2, también se parece a la UMH en MASTER2 y MASTER5. Por otra parte, las dos universiades más grandes UPV y UV parecen ser casi completamente opuestas.

3.4.3. Coordenadas paralelas

Este es otro de los gráficos que sirven para comparar muchas variables juntas y ver relaciones entre ellas.

En este caso, en el eje de las X se colocan las variables y en el eje de las Y tenemos los diferentes valores que pueden tomar. En este caso sólo lo harán del 0 al 1 porque están estandarizadas en ese intervalo. La función *parcoord* [20] del paquete *MASS* [21] de R dibujará

cada observación según el valor que tome cada variable en ella y se visualizarán todas juntas.

```
parcoord(variables)
```

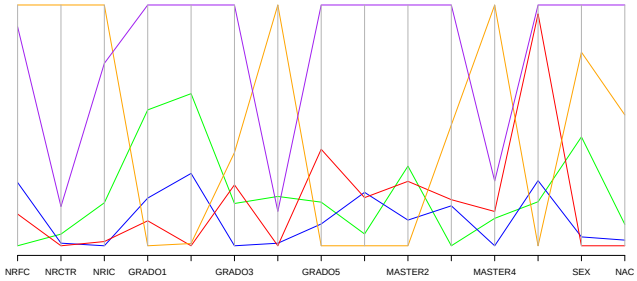


Figura 3.9: Gráfico de coordenadas paralelas.

Cada color del gráfico de la figura 3.9 se corresponde con una observación diferente. La UA toma el color azul, la UJI el color verde, el rojo se corresponde a la UMH y el naranja y el morado a la UPV y UV respectivamente.

Con este gráfico terminamos de confirmar lo que veníamos analizando en los dos anteriores. En aquellas variables en las que la UPV toma sus valores más altos, la UV por el contrario toma los más bajos excepto en la variable SEXENIOS. En cuanto a la UJI parece que sigue la misma tendencia que la UA aunque con valores ligeramente menores excepto en NRFC, MASTER1, MASTER3 y MASTER5 donde toma valores algo mayores. Por otra parte la UMH podría seguir un recorrido parecido a la UV aunque en mucha menor escala, alcanzándola sólo en la variable MASTER5.

Capítulo 4

Conclusiones

Las prácticas realizadas en este proyecto se han llevado a cabo en el Gabinete de Planificación y Prospectiva Tecnológica de la UJI con el objetivo de conocer el PPF del SUPV, realizar diferentes simulaciones y analizar las variables que intervienen en él.

Durante la estancia en prácticas asociada a este proyecto se ha comenzado realizando una búsqueda en diferentes sitios web para conseguir los datos necesarios de las universidades del SUPV que intervienen en el PPF y tratarlos adecuadamente. Con estos datos se han realizado diferentes simulaciones de la aplicación del PPF a las 5 universidades públicas de la Comunidad Valenciana y además se han creado diferentes informes donde podemos ver las diferencias de subvenciones recibidas entre ellas y evaluar la subvención recibida por la UJI a través de diferentes años. Posteriormente, se ha realizado otra búsqueda en internet con el objetivo de estudiar otros planes de financiación diferentes. A partir de ellos se han simulado cambios en el PPF original del SUPV para intentar mejorar el beneficio asociado a la UJI.

Por otra parte, en la memoria del TFG, se han estudiado los fundamentos teóricos que se corresponden con las diferentes técnicas estadísticas realizadas para analizar el PPF. Como son los diferentes algoritmos de imputación proporcionados por el software R, el análisis de componentes principales; el cual nos ha permitido realizar una comparativa de similitud entre las universidades a partir de diferentes variables. Y, por último, la creación de diferentes métodos gráficos multivariantes para poder representar los datos con los que se ha trabajado en este proyecto y estudiar los resultados obtenidos.

Bibliografía

[1] Plan plurianual de financiación de las universidades públicas de la comunidad valenciana. https://gerencia.ua.es/es/documentos/documentos433/plan-plurianual-de-financiacion-2010-2017.pdf.

[2] Sistema de información de las universidades valencianas públicas. http://www.siuvp.es/es/.

[3] Base de datos del ivie. https://www.ivie.es/es_ES/bases-de-datos/.

[4] La universidad española en cifras. http://www.crue.org/SitePages/La-Universidad-Espa~nola-en-Cifras.aspx.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[6] The r project for statistical computing. https://www.r-project.org/.

[7] Husson and Josse. Package missmda. https://cran.r-project.org/web/packages/missMDA/missMDA.pdf.

[8] Julie Josse and François Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2), 2012.

[9] Daniel J. Stekhoven. missforest function. https://www.rdocumentation.org/packages/missForest/versions/1.4/topics/missForest.

[10] Lior Rokach and Oded Z Maimon. *Data mining with decision trees: theory and applications*, volume 69. World scientific, 2008.

[11] Daniel Peña. *Análisis de datos multivariantes*. 2002.

[12] Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[13] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 2011.

[14] Daniel Zelterman. *Applied multivariate statistics with R.* Springer, 2015.

[15] Hans Peter Wolf. function function. `https://www.rdocumentation.org/packages/aplpack/versions/1.3.3/topics/faces`.

[16] Hans Peter Wolf. aplpack package. `https://cran.r-project.org/web/packages/aplpack/index.html`.

[17] Gráficos estadísticos con r. `https://cran.r-project.org/doc/contrib/grafi3.pdf`.

[18] R-core. stars function. `https://www.rdocumentation.org/packages/graphics/versions/3.6.1/topics/stars`.

[19] R-core. graphics package. `https://www.rdocumentation.org/packages/graphics/versions/3.6.1`.

[20] Brian Ripley. parcoord function. `https://www.rdocumentation.org/packages/MASS/versions/7.3-51.4/topics/parcoord`.

[21] Brian Ripley. Mass package. `https://cran.r-project.org/web/packages/MASS/MASS.pdf`.

[22] Diego C Araneo. Introducción al análisis de componentes principales. 2008.

[23] Andrés Sánchez Mangas. Análisis de componentes principales: versiones dispersas y robustas al ruido impulsivo. Master's thesis, 2012.

[24] José Javier Martínez Fernández de las Heras et al. La descomposición en valores singulares (svd) y algunas de sus aplicaciones. *Gaceta de la Real Sociedad Matematica Española*, 8(3):796–810, 2005.

[25] Descomposición en valores singulares. `http://www.mate.unlp.edu.ar/practicas/70_18_0911201012951`.

[26] Javier Sanchís Saez. *GPC mediante descomposición en valores singulares (SVD). Análisis de componentes principales (PCA) y criterios de selección.* PhD thesis, 2009.

[27] Torsten Hothorn and Brian S Everitt. *A handbook of statistical analyses using R.* Chapman and Hall/CRC, 2014.

[28] Michael J Crawley. *The R book.* John Wiley & Sons, 2012.

Anexo A

Anexo I

Figura A.1: Informe Evolución UJI

Figura A.2: Informe Evolución UJI

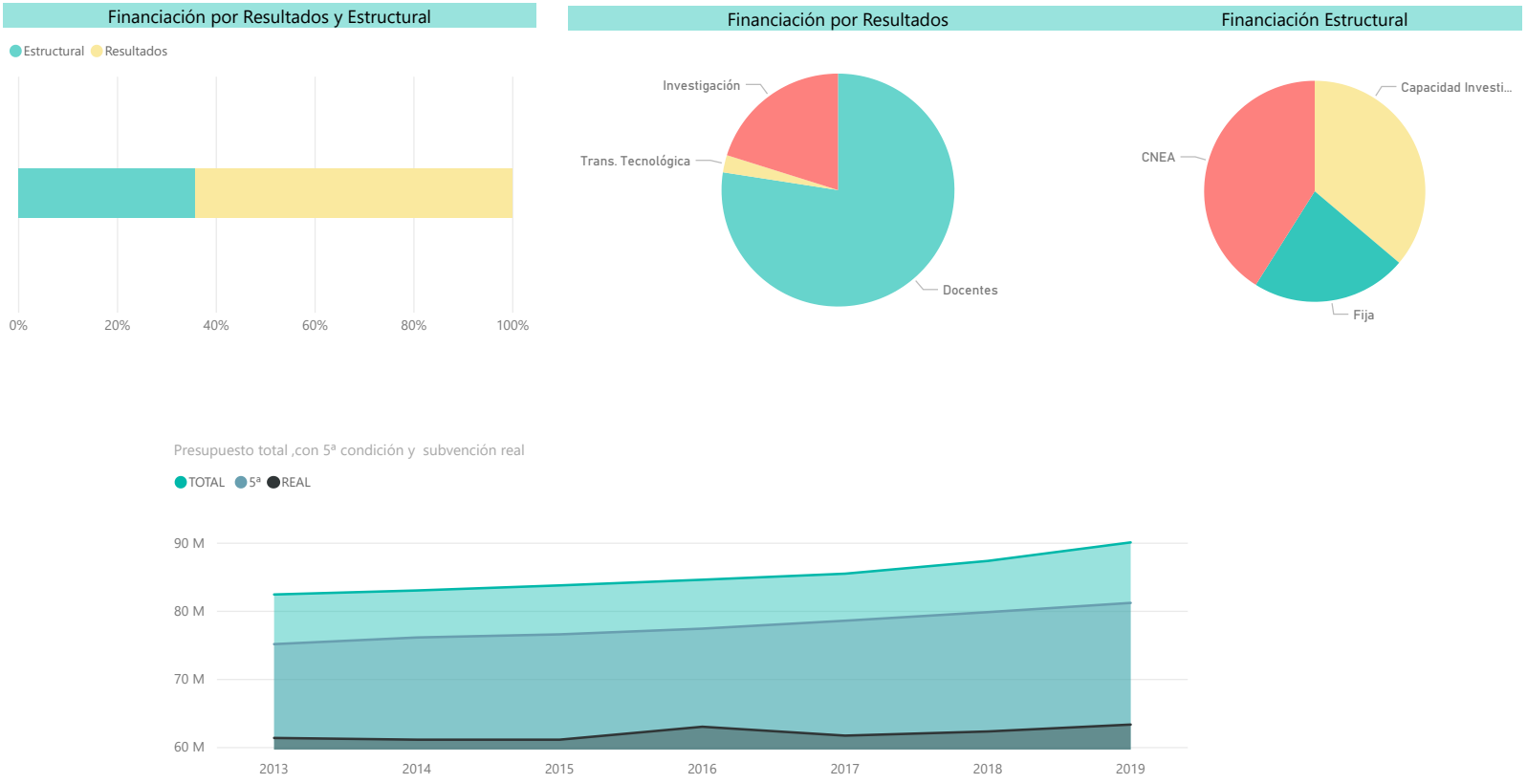


Figura A.3: Informe Evolución UJI

Figura A.4: Informe comparación SUPV

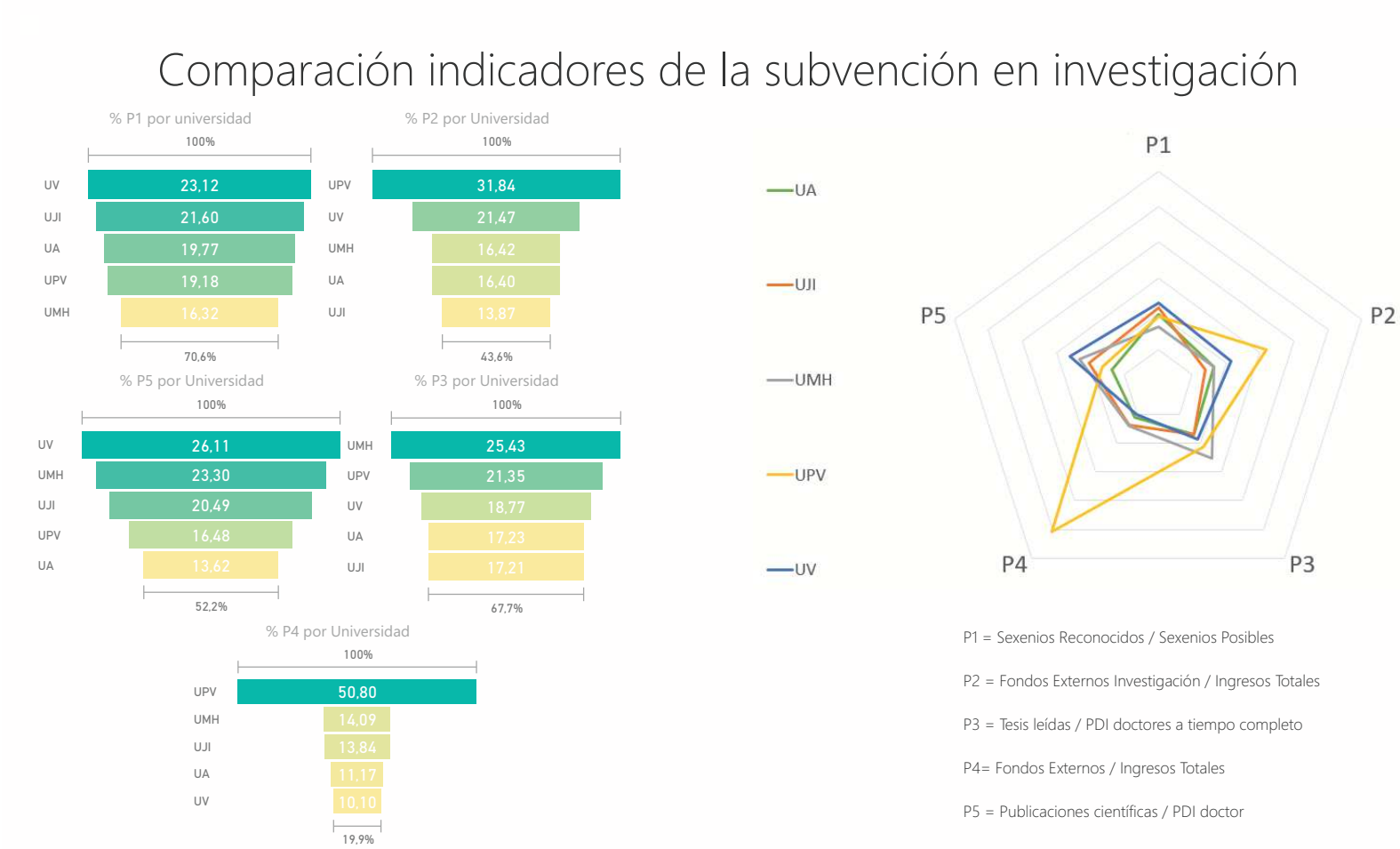


Figura A.5: Informe Investigación Madrid en PPF SUPV

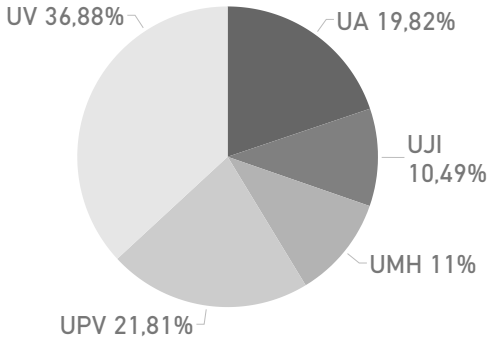


Figura A.6: % SRD según experimentalidades Madrid

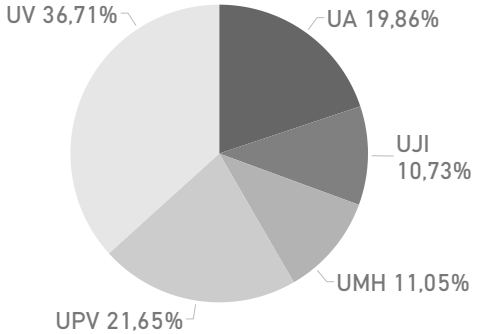


Figura A.7: % SRD según experimentalidades SUPV

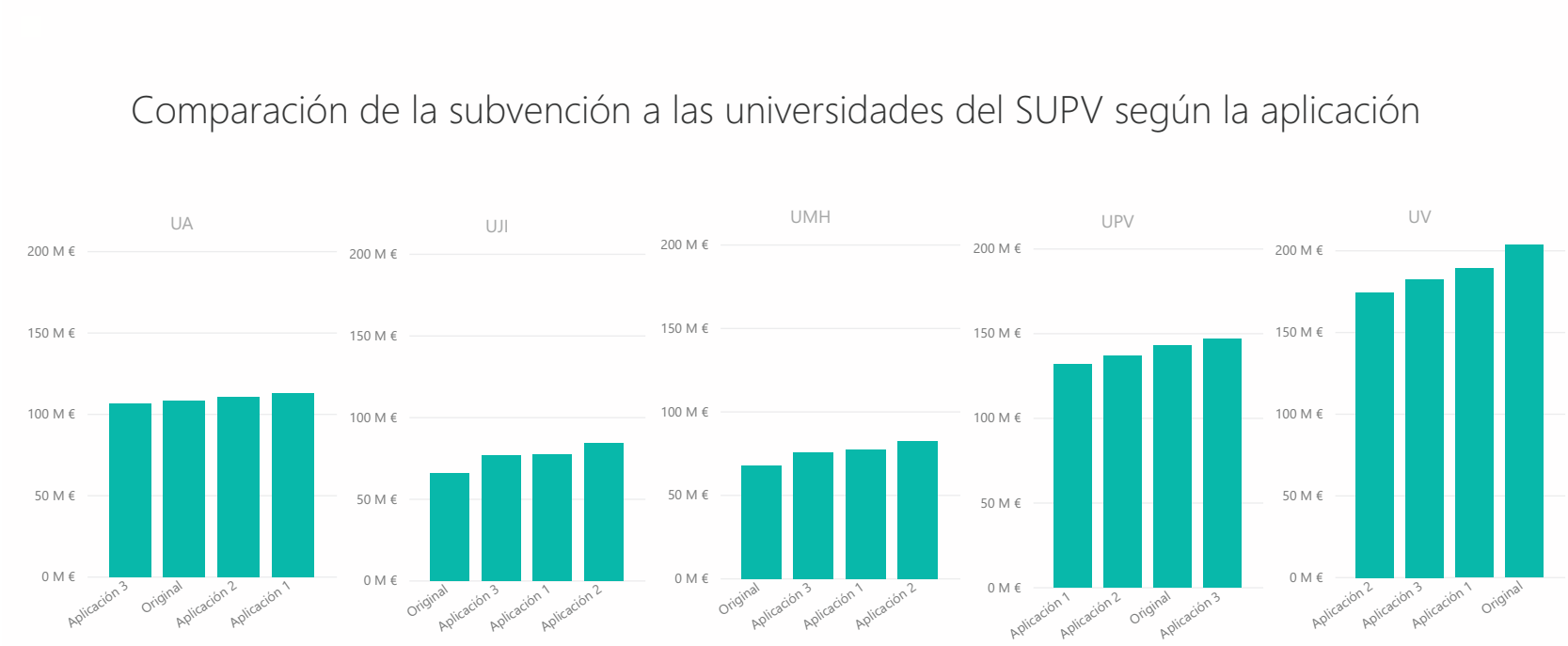
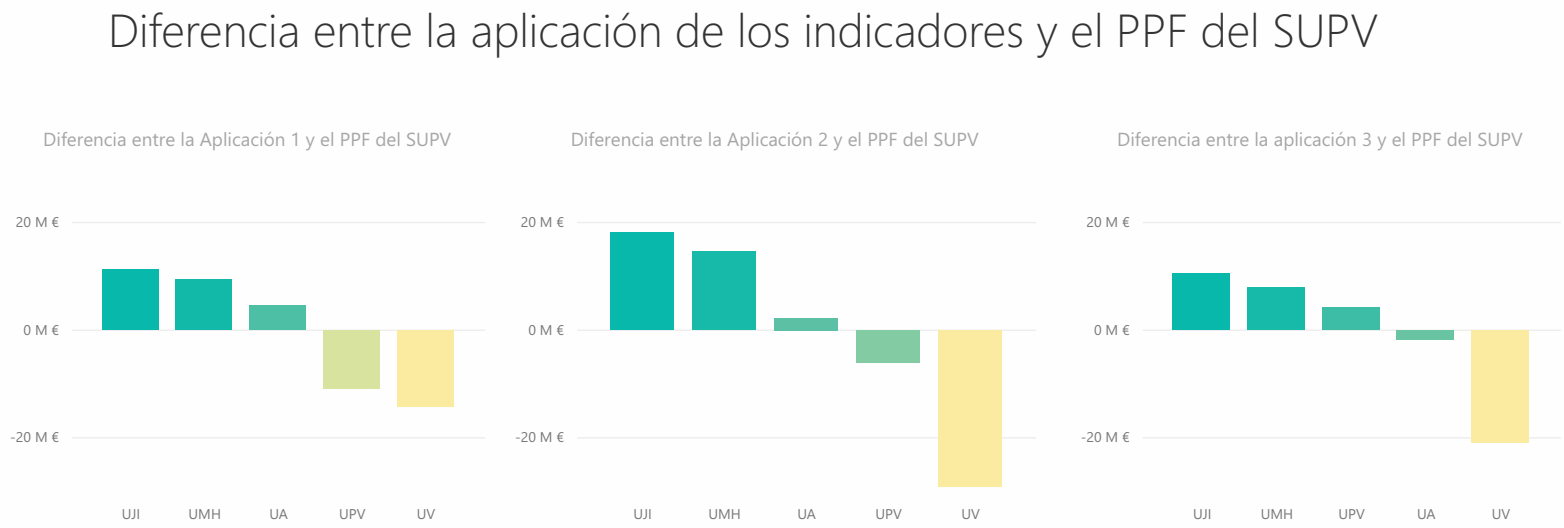


Figura A.8: Informe Investigación Madrid en PPF SUPV



Podemos observar que las universidades más favorecidas son las universidades más pequeñas, como la UJI y la UMH. Mientras que la universidad que se ve más afectada negativamente es la UV.

Figura A.9: Informe Investigación Madrid en PPF SUPV

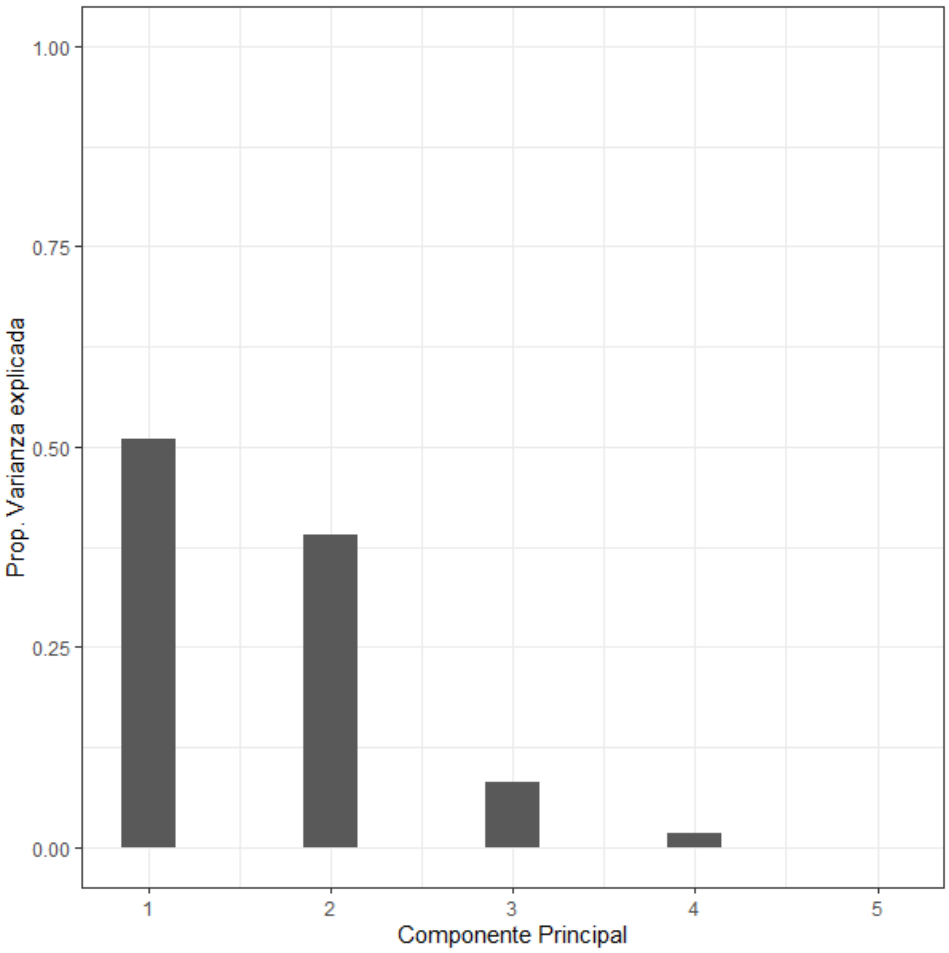


Figura A.10: Proporción de varianza por cada CP

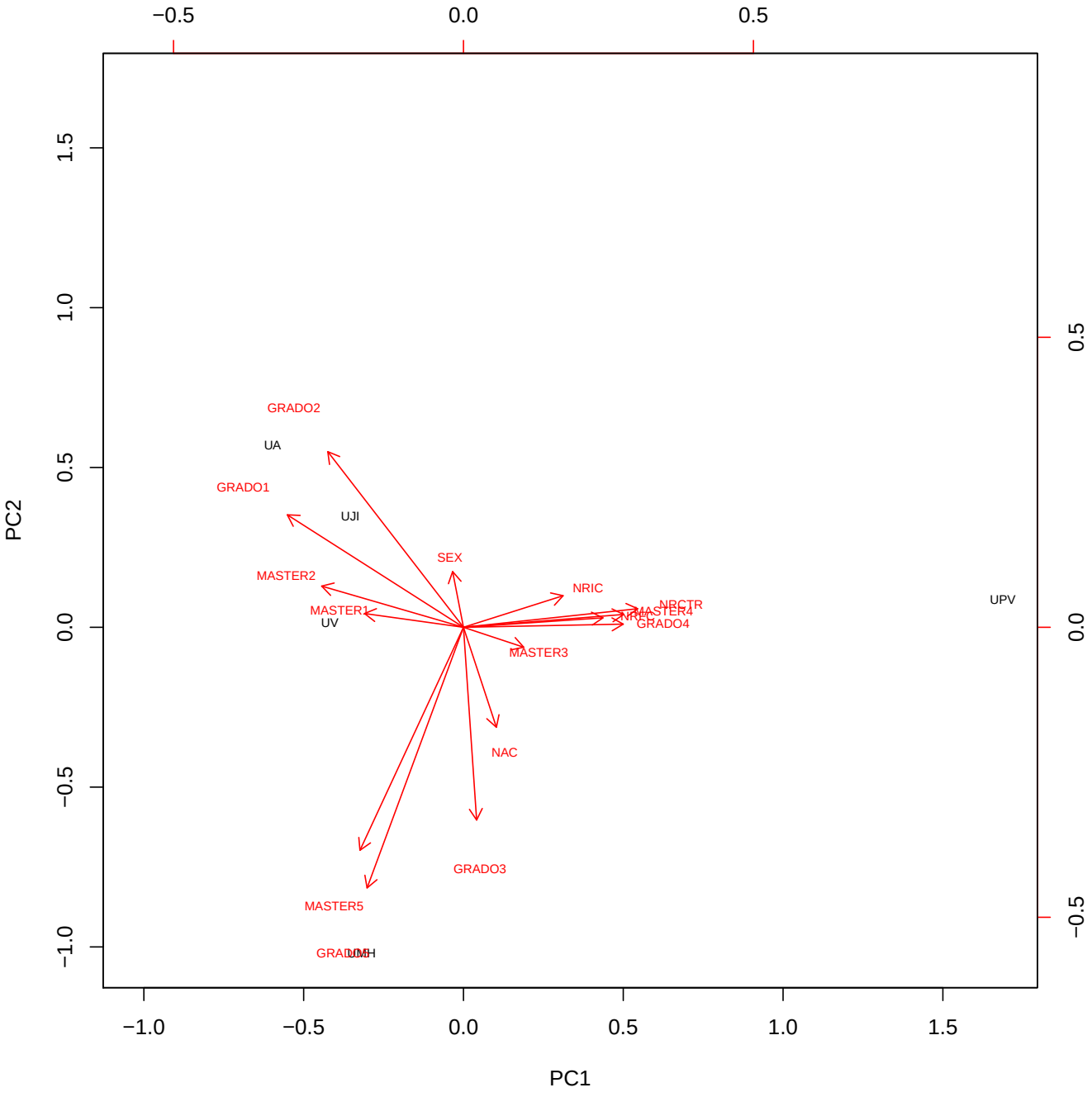


Figura A.11: Representación de las dos primeras CP

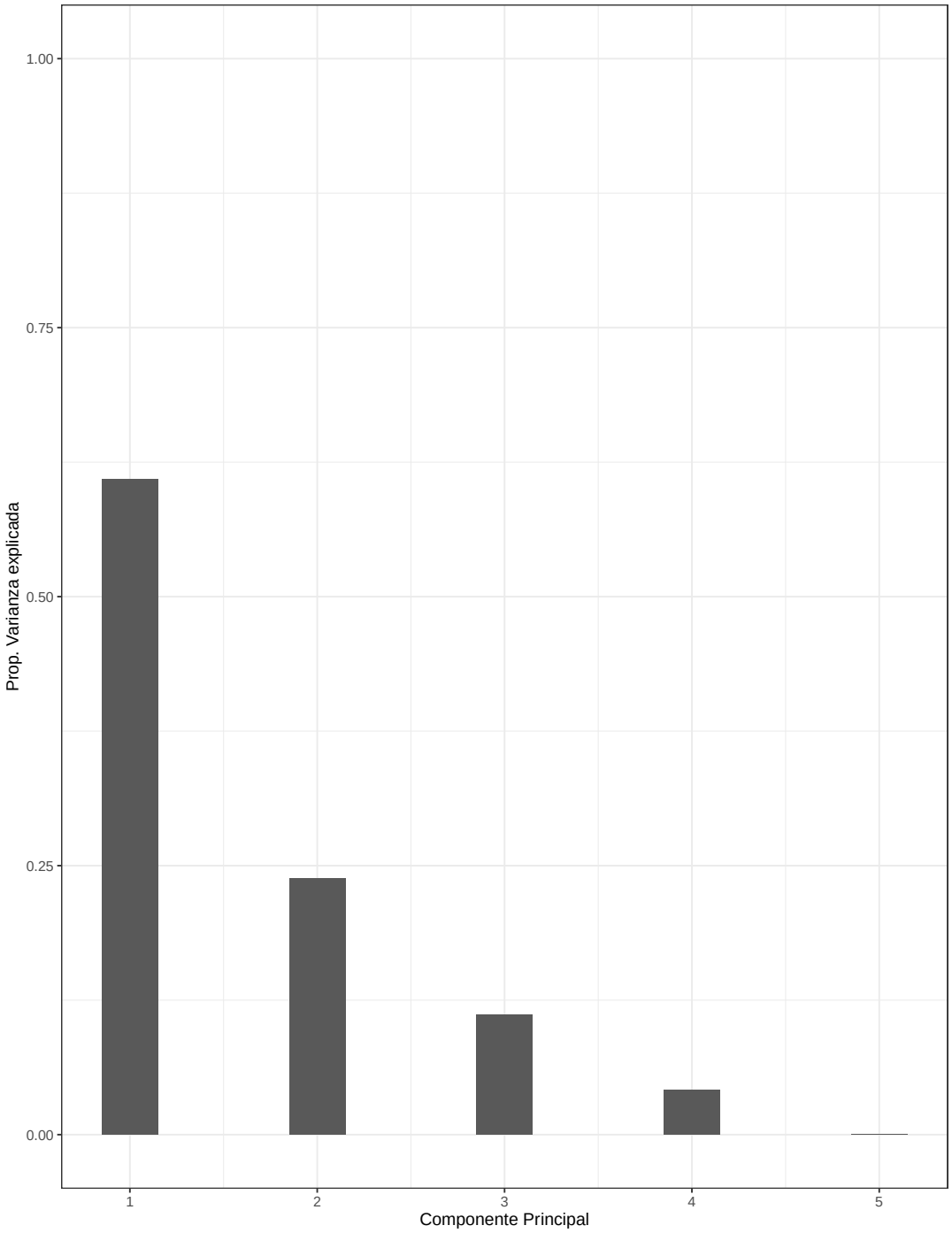


Figura A.12: Proporción de varianza de las dos primeras CP

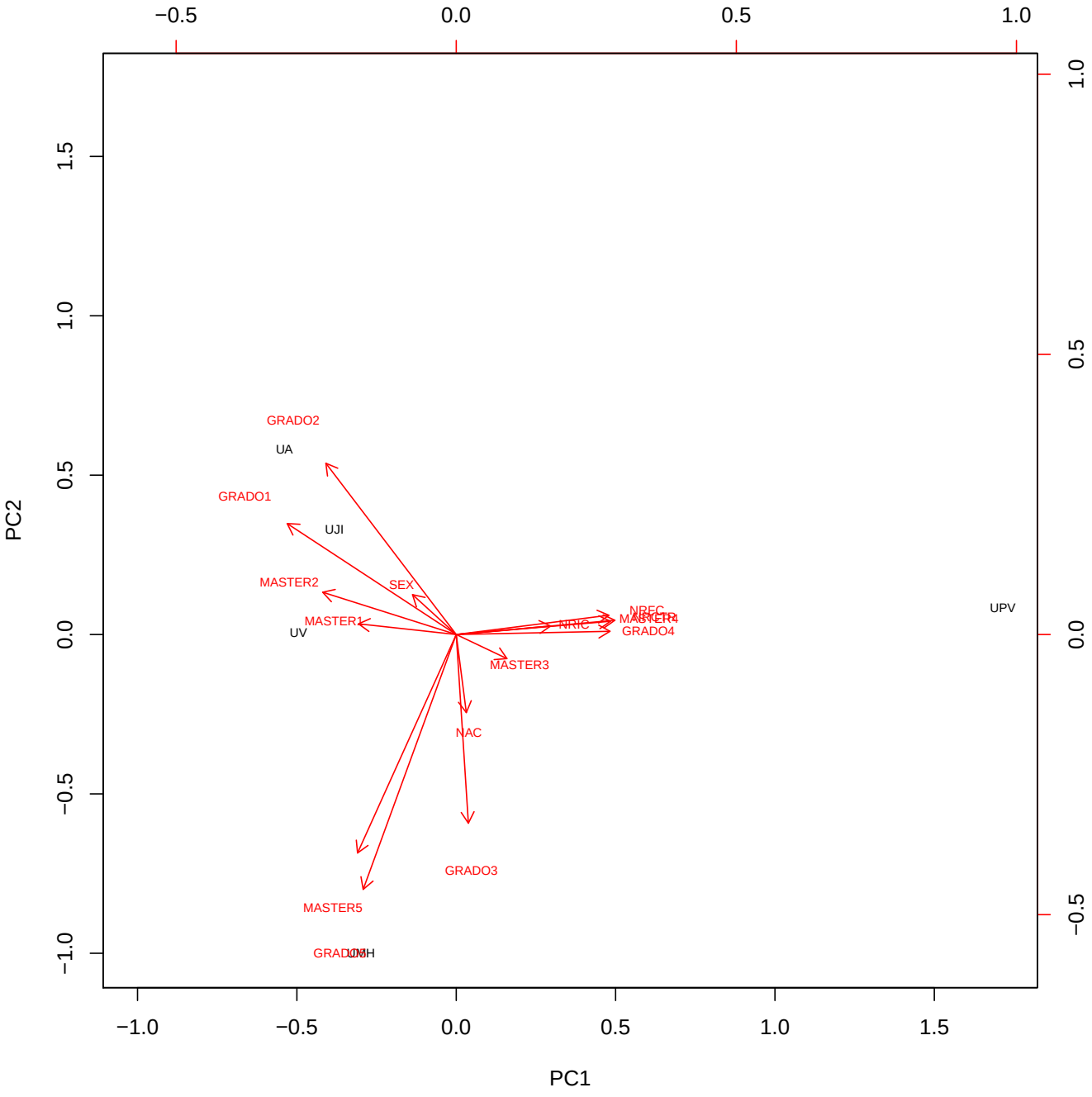


Figura A.13: Representación dos primeras CP