UNIVERSIDAD DE MURCIA
Facultad de Matemáticas

TRABAJO FIN DE GRADO

PCA-Gapminder, datos multivariantes en una nueva visión.

Realizado por:
Inés Amorós Fernández

Dirigido por:
Jorge Luis Navarro Camacho

Junio 2017

DECLARACIÓN DE ORIGINALIDAD

Yo, *Inés Amorós Fernández*, autora del TFG "PCA-Gapminder, datos multivariantes en una nueva visión", bajo la tutela del profesor Jorge Luis Navarro Camacho, declaro que este trabajo es original en el sentido de que he puesto el mayor empeño en citar debidamente todas las fuentes empleadas para su elaboración.

En Murcia, a 13 de junio de 2017.

Nota: En la Secretaría de la Facultad de Matemáticas se ha presentado una copia firmada de esta declaración.

"No podemos enseñar nada a nadie. Tan sólo podemos ayudar a que descubran por sí mismos."

Galileo Galilei.

Resumen

Pocos términos admiten aplicarles en tan alto grado el concepto de transversalidad como la palabra estadística. Eso sí, transversalidad caótica que unas veces se mantiene dentro de los límites científicos y otras, desafortunadamente, ocupa espacios alejados del rigor exigible. No negaremos que para un matemático es irritante la expansión incontrolada, superficial, arbitraria o interesada de esta disciplina. Sin embargo, también es posible contemplar esta circunstancia desde el asombro, la avidez de indagar los porqués o como una oportunidad para cumplir un ineludible objetivo universitario: contribuir a la producción y divulgación científica.

Justificamos nuestro trabajo aludiendo a competencias básicas y específicas del TFG. Entre las primeras CB3 (Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes para emitir juicios que incluyan una reflexión sobre temas de índole social, científica o ética) y CB4 (Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado). Y respecto a las segundas: CG3 (Ser capaz de gestionar la información y el conocimiento en el ámbito de la Matemática, incluyendo saber utilizar como usuario las herramientas básicas en TIC), CE2 (Proponer, analizar, validar e interpretar modelos de situaciones reales sencillas, utilizando las herramientas matemáticas más adecuadas a los fines que se persigan) y CE3 (Utilizar aplicaciones informáticas de análisis estadístico, cálculo numérico y simbólico, visualización gráfica, optimización u otras para experimentar en Matemáticas y resolver problemas).

El término Estadística - [1]Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades - procede del alemán Statistik, derivado a su vez del italiano statista "hombre de Estado". La razón de esta etimología se explica así en la página del Instituto Nacional de Estadística (INE)

"...los gobiernos de las distintas naciones tenían la necesidad, por razones de organización, de conocer las características de su población para gestionar el pago de impuestos, el reclutamiento de soldados, el reparto de tierras o bienes, la prestación de servicios públicos..."

Pero la estadística ha ido evolucionando poniéndose al servicio de la sociedad. Se ha convertido en una herramienta esencial, una disciplina que ayuda a entender el mundo que nos rodea y un recurso básico para la toma de decisiones. Disponemos en la actualidad de una cantidad ingente de datos; relaciones interminables de tablas Excel, cientos de gráficos que nos ofrecen información puntual de todo lo imaginable y hasta de lo no imaginable. El reto en la actualidad no es obtener más datos, ni siquiera trasladar los datos a herramientas sencillas; el reto en la actualidad es conseguir interrelacionarlos y extraer información relevante de ellos.

[1]acepción 5. DRAE, vigesimotercera edición.

Nos referimos, por tanto, a un problema real, a una dificultad observable, que podemos enunciar como la dificultad de comprender procesos en los que intervienen muchas variables. En la mayoría de ocasiones todas las variables pedirán ser estudiadas simultáneamente con el fin de comprender por completo la estructura y las características clave de los datos. Surgen entonces las técnicas multivariantes, una de las cuales se aborda en este proyecto, el análisis de componentes principales (conocido como PCA, Principal Component Analysis). El objetivo de esta técnica, en sentido general, es exponer o extraer la señal que se esconde en los datos en presencia del ruido y descubrir qué es lo que el conjunto de datos nos muestra en medio de su aparente caos.

El instrumento que permite a los estadísticos hallar deducciones acerca de los datos estudiados es la capacidad de generar gráficas reveladoras. Esta capacidad se ha convertido en una especialidad dentro de la Estadística en los últimos años. Se han sugerido muchas maneras ingeniosas de representar datos utilizando animaciones y colores para la visualización. Dentro de estos destaca sobresalientemente la aplicación Gapminder.

Originariamente, se trata de un software que permite trabajar con hasta cinco variables, dos en los ejes, una en el tamaño de la burbuja, otra en el color y la última con el tiempo. Su autor, Hans Rosling, fallecido el pasado mes de febrero, quiso promover una visión del mundo basada en datos. Gapminder dio movimiento a millones de cifras que, olvidadas o sepultadas, acumulaban los gigantescos bancos de datos de las principales instituciones del planeta.

Inevitablemente, surgió la idea de aunar estadística multivariante, o más específicamente Análisis de Componentes Principales, y Gapminder. Así, uniendo PCA y Gapminder, proponemos la nueva herramienta gráfica PCA-Gapminder con el propósito de relacionar múltiples variables, conectándolas en conjunto, facilitando su comprensión y el desciframiento de las relaciones existentes. La programación de esta nueva herramienta se llevará a cabo mediante RStudio. La principal novedad que aporta PCA-Gapminder es que en los ejes no se representarán las variables, sino las componentes principales. Aunque en el capítulo primero se estudiará en profundidad, podemos adelantar que en líneas generales, estas componentes principales son nuevas variables formadas como una combinación lineal de las originales, y además sin correlación lineal entre ellas.

Contaremos, pues, con una aplicación capaz de representar cinco variables. Es incuestionable la gran utilidad que tiene este programa para la representación de muestras de datos con muchas variables. Reducirá la dimensionalidad de las variables mediante el PCA, representando esas componentes principales en los ejes; pero además quedarán todavía dos variables libres para representar con el color y el tamaño. Será interesante representar con ellas variables que se hayan perdido al hacer el PCA o alguna de tipo cualitativo. Finalmente se completará esta representación con una variable temporal, aportando un dinamismo que ayudará al público a tener una visión ágil y con perspectiva de la situación estudiada en un periodo de tiempo.

En el capítulo primero de este proyecto nos sumergiremos en toda la base teórica del Análisis de Componentes Principales. Se introducirán primero las ideas en las que se basa esta técnica desde distintos puntos de vista matemáticos, como el geométrico o estadístico; y descubriremos el nuevo concepto de componente principal, llegando finalmente al teorema más importante, precisamente el Teorema de las Componentes Principales, que demuestra su existencia y unicidad (bajo ciertas condiciones) y permite el cálculo de las componentes mediante la diagonalización de matrices. Por último, se comentarán los comandos que ofrece RStudio para el uso de esta técnica.

En el segundo capítulo nos centraremos en describir cómo surgió Gapminder de la mano de Hans Rosling y en nuestra extensión de dicha aplicación hasta convertirse en PCA-Gapminder, que permite que todas (o casi todas) las variables se puedan representar en un único gráfico dinámico que mantiene un porcentaje alto de la información contenida en todos los datos.

Finalmente, en el capítulo tercero, se han escogido dos casos de interés para analizar utilizando la nueva aplicación. Respecto a este último aspecto, es conveniente apuntar algunas precisiones. Una vez acotados los puntos preliminares del proyecto, llegó el momento de seleccionar el campo de aplicación. Las opciones de trabajo eran tan abundantes y variadas que costó tomar una decisión. ¿Qué escogíamos? ¿Qué descartábamos? Disciplinas científicas o humanísticas, saberes clásicos o de última tecnología, hechos de actualidad o acontecimientos revisados por el filtro de los siglos... Tras algunos titubeos nacidos del interés que suscitaban numerosos temas, optamos por dos ámbitos, uno interno (nuestra Facultad) y otro externo (la situación de las diferentes Comunidades Autónomas).

Ante los cuantiosos gráficos, el observador podrá esbozar muchas interpretaciones. Están al alcance de cualquiera, solo con un poco de tiempo y atención. Ese es el principal mérito de la aplicación. Sin embargo, para responder con honestidad a las exigencias de este trabajo fin de grado, hemos de asumir que el bloque de interpretación no se ha desarrollado completamente. Hemos realizado un esfuerzo en la deducción de conclusiones, pero somos conscientes de que es un resultado modesto. Para extraer el máximo rendimiento deductivo a la aplicación que presentamos precisaríamos de información contextual más detallada –en el caso 1- o de mayor rigor sociológico –en el caso 2- No obstante, no es este el objetivo de nuestro proyecto y la información que proporcionamos es, en nuestra opinión, un adecuado punto de partida para futuras investigaciones y, principalmente, sirve para mostrar la aplicación de la nueva herramienta, que es justamente lo que queríamos hacer.

Respecto al análisis que presentamos, hemos fijado nuestra atención en algunos aspectos que pueden resultar de interés, bien por su cercanía (nuestra Facultad, nuestra Región), bien por su trascendencia para el futuro (nivel de formación de los ciudadanos, estabilidad laboral). Se trata de indicadores que, en su ámbito de influencia, son valiosos pero que, contemplados aisladamente, sin anclajes contextuales o desgajados de un todo, pueden apartarse de la necesaria objetividad que un trabajo riguroso precisa. De ahí que no se enuncien conclusiones definitivas en los apartados finales. Este aspecto le correspondería a un proyecto sociológico que, con métodos científicos cualitativos y/o comparativos, abordase las causas y consecuencias de la realidad descrita.

Por el contrario, este proyecto de trabajo sí avanza en otros objetivos innovadores. En primer lugar, lograr presentar una realidad n-dimensional convenientemente adaptada a la percepción humana, que es bidimensional. En segundo lugar, explorar y aprovechar las posibilidades de un instrumento como PCA-Gapminder. Y finalmente, presentar este instrumento como una herramienta valiosa, útil para explicar macrotendencias sociales, culturales o económicas, para evitar ideas preconcebidas, para entender por qué un indicador parece estancado y otro avanza con rapidez y, en último extremo, incluso para predecir la evolución futura.

Abstract

There are only a few terms that admit applying to such a high level the concept of transversality as the word statistic does. This is, chaotic transversality, which sometimes remains within the scientific limits and others, unfortunately, occupies spaces far from the exacting rigor. We will not deny that for a mathematician it is irritant the uncontrolled, superficial, arbitrary or self-serving expansion of this discipline. However, it is also possible to consider this fact from the astonishment, the eagerness to investigate the reasons or as an opportunity to fulfill an unavoidable university aim: to contribute to the scientific progress and dissemination.

We justify our work by referring to the basic and specific competencies of the TFG. Among the first ones, CB3, (students should develop the ability to gather and interpret relevant information to make judgments that include a reflection on social, scientific or ethical issues) and CB4 (students should be able to transmit information, ideas, problems and solutions to both specialized and non-specialized public). Regarding the second ones: CG3 (being able to manage information and knowledge in the field of mathematics, including knowing how to use basic ICT tools as a user), CE2 (proposing, analyzing, validating and interpreting models of simple real situations by using the mathematical tools most suited to the desired purposes) and CE3 (To use computer applications of statistical analysis, numerical and symbolic calculation, graphical visualization, optimization or others to experiment in Mathematics and to solve problems).

The term Statistical - [2]Branch of mathematics uses large sets of numerical data to obtain inferences based on the calculation of probabilities - comes from the German Statistik, which derived from the Italian statist "man of State". The reason for this etymology is explained in the page of the National Institute of Statistics of Spain (INE).

"... the governments of the different nations had the need, for organizational reasons, to know the characteristics of their population to manage the payment of taxes, the recruitment of soldiers, the distribution of land or property, the provision of public services, ..."

But statistics have evolved by serving society. It has become an essential tool, a discipline that helps to understand the world around us and a basic resource for decision making. We currently have an enormous amount of information; endless relationships of Excel tables as well as hundreds of graphs that offer us timely information of everything we can imagine and even of the unimaginable. Nowadays, the challenge is not to obtain more data or to transfer the information to simple tools; today, the challenge is to interrelate and to obtain information from them.

Therefore, we are referring to a real problem; an observable challenge which can be enunciated as the difficulty of understanding processes in which many variables intervene. In the majority of cases, all variables will be asked to be studied simultaneously in order to fully un-

[2]meaning 5. DRAE, twenty-third edition

derstand the structure and key characteristics of the data. As a result, multivariate techniques arise, one of which is addressed in this project, the analysis of principal components (known as PCA, Principal Component Analysis). Overall, the aim of this technique is to expose or extract the signal hidden in the data in the presence of noise as well as to discover what the data set shows us in the middle of its apparent chaos.

The instrument that allows statisticians to find deductions about the data studied is the ability to generate revealing graphs. This capacity has become a specialty within Statistics in recent years. Many ingenious ways of representing data using animations and colors for visualization have been suggested. Among these, the Gapminder application stands out.

Originally, it is a software that allows working with five variables: two in the axes, one in the size of the bubble, another in the color and the last one with the time. Its author, Hans Rosling, who died last February, wanted to promote a worldview based on data. Gapminder gave movement to millions of figures that, forgotten or buried, accumulated the gigantic data banks of the main institutions of the world.

Inevitably, the idea of combining multivariate statistics, or more specifically Principal Component Analysis, and Gapminder arose. PCA and Gapminder, the new PCA-Gapminder graphical tool is proposed here with the purpose of relating multiple variables, connecting them together, facilitating their understanding and deciphering existing relationships. The programming of this new tool will be carried out by RStudio. The main new element of PCA-Gapminder is that it will not be the axes which represent the variables but the main components. Although in the first chapter it will be studied in depth, we can say that in general, these main components are new variables formed as a linear combination of the originals, and also without linear correlation between them.

Accordingly, we will have an application capable of representing five variables. Then, it is unquestionably the great utility that this program has for the representation of large amounts of data. It will reduce the dimensionality of the variables through the PCA, being represented those main components in the axes; but there are still two free variables to represent with color and size. It will be interesting to represent with them those variables that have been lost in doing the PCA. Finally, this representation will be completed with a temporary variable, contributing a dynamism that will help to have a perspective view of the situation studied in a period of time.

In the first chapter of this project, we will immerse ourselves in the whole theoretical basis of the Analysis of Principal Components. First of all, we will introduce the ideas on which this technique is based from different mathematical points of view, such as the geometrical or statistical one. We also will discover the new concept of principal component, culminating with the most important theorem. Finally, some commands offered by RStudio will be exposed for the use of this technique.

In the second chapter, we will focus on describing how Gapminder arose from the hand of Hans Rosling and on the extent of that application to become PCA-Gapminder. Finally, two cases of interest to analyze by using the new application have been studied in chapter three.

Regarding the last aspect - the cases that exemplify the application developed - it is convenient to point out some details. Once the preliminary points of the project were delimited, it was time to select the field of application. The work choices were so plentiful and varied that it took a long time to choose between them. What should be chosen? What should be selected? Scientific or humanistic disciplines? Classical or state-of-the-art knowledge? Current events or events reviewed by the screen of the years . . . ? After some hesitation stemming from the interest aroused by many issues, we selected two domains, one internal (our Faculty) and another external (the situation of the different Autonomous Communities).

In the light of plenty graphs, the observer could sketch many interpretations. They are achievable by anyone with a little time and attention. That is the main credit of the application. However, to respond honestly to the demands of this end-of-grade paper, we must assume that the block of interpretation has not been fully developed. We have made an effort to deduce conclusions, but we are aware that this is a modest result. To extract the maximum deductive performance to the application that we present we would need more detailed contextual information - in the case 1- or information of more sociological rigor - in the case 2-. Nevertheless, this is not the main purpose of the project and the provided information is, in our opinion, an adequate starting point for future research and enough to show how to apply the new technique.

Regarding the analysis we present, we have focused our attention on some aspects that may be of interest because of its proximity (near our Faculty and our Region) or because of its importance for the future (training level of citizens, job stability). These are indicators that, in their sphere of influence are valuable but viewed in isolation and out of place or detached from a context, they can deviate from the necessary objectivity that a rigorous work needs. For this reason, no definitive conclusions are stated in the final sections. This aspect would correspond to a sociological project that, with qualitative or comparative scientific methods, would address the causes and consequences of the described reality.

On the contrary, this project advances in other innovative aims. First, it achieves an n-dimensional reality conveniently adapted to human perception, which is two-dimensional. Secondly, it explores and takes advantage of the possibilities of an instrument such as PCA Gapminder. And finally, the project presents this instrument as a valuable and useful tool for the expansion of social, cultural or economic macro trends as well as to avoid preconceived ideas, to understand why one indicator seems stagnant while another advances rapidly and, ultimately, even to predict a future evolution.

Índice general

Capítulo 1

Marco teórico. Componentes principales

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad. Si es posible describir con precisión los valores de p variables por un pequeño subconjunto r < p de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información[1].

El análisis de componentes principales tiene este objetivo: busca combinaciones lineales que puedan ser usadas para resumir los datos, perdiendo la mínima información en el proceso. Es decir, dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un menor número de variables construidas como combinaciones lineales de las originales. Por ejemplo, con variables con alta dependencia es frecuente que un pequeño número de nuevas variables (menos del 20 por ciento de las originales) expliquen la mayor parte (más del 80 por ciento de la variablidad original). La utilidad de esta técnica es doble:

- Permite representar óptimamente en un espacio de dimensión pequeña observaciones de un espacio general p-dimensional. En este sentido, componentes principales es el primer paso para identificar las posibles variables latentes, o no observadas que generan los datos.
- Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

1.1. Introducción histórica

El descubrimiento del PCA se le suele atribuir por completo, en el año 1933, a Harold Hotelling; pero esto no es del todo cierto. Hubo muchas aportaciones anteriores. Los primeros estudios sobre relación entre variables se deben a Adolfo Quetlet, que aplicó por primera vez métodos probabilísticos a las ciencias sociales. Ahora bien, fue Francis Galton el primero en introducir modestos inicios en el Análisis de Componentes Principales. Advirtió la necesidad de eliminar información redundante en un conjunto de variables aleatorias.

Más adelante, en 1901, McDonald hizo un estudio con 7 variables y 3000 individuos, hallando los resultados de una matriz de correlaciones, con el fin de encontrar algún "índice" que resumiese la información contenida en los datos. Posteriormente, Karl Pearson obtuvo el estimador del coeficiente de correlación en muestras, y afrontó el problema de la determinación de si dos

[1]Los resultados desarrollados en este capítulo se han obtenido de las fuentes [3] y [4] citadas en la Bibliografía.

grupos de personas, de los que se conocen sus medidas físicas, pertenecen a la misma raza.

Este problema intrigó a Harold Hotelling. Y fue en este momento, en 1929, cuando decide viajar a la estación de investigación agrícola de Rothamsted, en Reino Unido. Allí trabaja con el reputado R.A. Fisher. En 1933, a su vuelta a la Universidad de Columbia, un profesor le plantea a Hotelling el problema de encontrar los factores capaces de interpretar los resultados obtenidos en un test de inteligencia a un conjunto de personas. Fue en este año cuando Hotelling descubrió las componentes principales, unos indicadores capaces de sintetizar de una manera óptima en un conjunto grande de variables.

1.2. Planteamiento del problema

Supongamos que se dispone de los valores de k-variables en n elementos de una población dispuestos en una matriz M de dimensiones n x k, donde las columnas contienen las variables y las filas los elementos. Supondremos que previamente hemos restado a cada variable su media, de manera que las variables de la matriz M tienen media cero y su matriz de covarianzas vendrá dada por 1/n M'M. El problema que se desea resolver es encontrar un espacio de dimensión más reducida que represente adecuadamente los datos. Puede abordarse desde tres perspectivas equivalentes.

i) Enfoque descriptivo:

Se desea encontrar un subespacio de dimensión menor que k tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible. Veamos cómo convertir esta noción intuitiva en un criterio matemático operativo. Consideremos primero un subespacio de dimensión uno, una recta. Se desea que las proyecciones de los puntos sobre esta recta mantengan, lo máximo posible, sus posiciones relativas. Para concretar, consideremos el caso de dos dimensiones (k=2). La figura indica el diagrama de dispersión y una recta que, intuitivamente, proporciona un buen resumen de los datos, ya que la recta pasa cerca de todos los puntos y las distancias entre ellos se mantienen aproximadamente en su proyección sobre la recta.

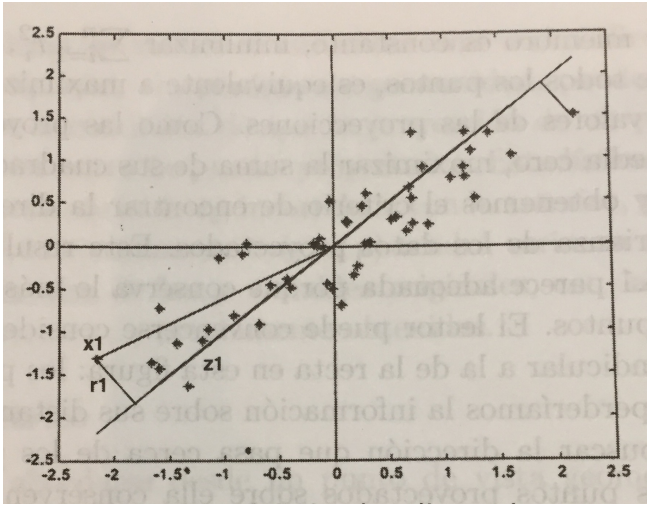


Figura 1.1: Ejemplo de recta que minimiza las distancias ortogonales de los puntos a ella.

La condición de que la recta pase cerca de la mayoría de los puntos puede concretarse exigiendo que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles. En consecuencia, si consideramos un punto x_i y una dirección $a = (a_1, \ldots, a_k)'$, definida por un vector a de norma unidad, la proyección del punto x_i sobre esta dirección es el escalar:

$$z_i = a_1 x_{i1} + ... + a_k x_{ip} = a' x_i \tag{1.1}$$

y el vector que representa esta proyección será $z_i a$. Llamando r_i a la distancia entre el punto x_i, y su proyección sobre la dirección a, este criterio implica:

$$\min \quad \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} |x_i - z_i a|^2 \tag{1.2}$$

donde $|u|$ es la norma euclídea o módulo del vector u.

La Figura 1.1 muestra que al proyectar cada punto sobre la recta se forma un triángulo rectángulo donde la hipotenusa es la distancia del punto al origen, $(x_i' x_i)^{1/2}$, y los catetos la proyección del punto sobre la recta (z_i) y la distancia entre el punto y su proyección (r_i). Por el teorema de Pitágoras, podemos escribir:

$$x_i' x_i = z_i^2 + r_i^2, \tag{1.3}$$

y sumando esta expresión para todos los puntos, se obtiene:

$$\sum_{i=1}^{n} x_i' x_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2, \tag{1.4}$$

Como el primer miembro es constante, minimizar $\sum_{i=1}^{n} r_i^2$, la suma de las distancias a la recta de todos los puntos, es equivalente a maximizar $\sum_{i=1}^{n} z_i^2$, la suma al cuadrado de los valores de las proyecciones. Como las proyecciones z_i son, por (1.1) variables de media cero, maximizar la suma de sus cuadrados equivale a maximizar su varianza, y obtenemos el criterio de encontrar la dirección de proyección que maximice la varianza de los datos proyectados. Este resultado es intuitivo: la recta de la Figura 1.1 parece adecuada porque conserva lo más posible la variabilidad original de los puntos.

ii) Enfoque estadístico:

Representar puntos k dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las k variables originales por una nueva variable, z_1, que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión. Esto no será posible si la nueva variable toma un valor semejante en todos los elementos, y, se demuestra, que la condición para que podamos aproximar con la mínima pérdida de información los datos observados, es utilizar las variables de máxima variabilidad.

Volviendo a la Figura 1.1 se observa que la variable escalar obtenida al proyectar los puntos sobre la recta sirve para prever bien el conjunto de los datos. La recta indicada en la figura no es la línea de regresión de ninguna de las variables con respecto a la otra, que se obtienen minimizando las distancias verticales u horizontales, sino la que minimiza las distancias ortogonales

o entre los puntos y la recta y se encuentra entre ambas rectas de regresión.

Este enfoque puede extenderse para obtener el mejor subespacio resumen de los datos de dimensión 2. Para ello, calcularemos el plano que mejor aproxima a los puntos. El problema se reduce a encontrar una nueva dirección definida por un vector unitario, B, que, sin pérdida de generalidad, puede tomarse ortogonal a a, y que verifique la condición de que la proyección de un punto sobre este eje maximice las distancias entre los puntos proyectados. Estadísticamente esto equivale a encontrar una segunda variable z_2, incorrelada con la anterior, y que tenga varianza máxima. En general, la componente z_r $(r < k)$ tendrá varianza máxima entre todas las combinaciones lineales de las k variables originales, con la condición de estar incorrelada con las $z_1, \ldots, z_{r-1}$ previamente obtenidas.

iii) Enfoque geométrico:

El problema puede abordarse desde un punto de vista geométrico con el mismo resultado final. Si consideramos la nube de puntos de la Figura 1.1 vemos que los puntos se sitúan siguiendo una elipse y podemos describirlos por su proyección en la dirección del eje mayor de la elipse. Puede demostrarse que este eje es la recta que minimiza las distancias ortogonales, con lo que volvemos al problema que ya hemos resuelto. En varias dimensiones, tendremos elipsoides, y la mejor aproximación a los datos es la proporcionada por su proyección sobre el eje mayor del elipsoide.

Intuitivamente, la mejor aproximación en dos dimensiones es la proyección sobre el plano de los dos ejes mayores del elipsoide y así sucesivamente. Considerar los ejes del elipsoide como nuevas variables originales supone pasar de variables correladas a variables ortogonales o incorreladas como veremos a continuación.

1.3. Cálculo teórico de las componentes principales

Supongamos que $X = (X_1, \ldots, X_k)'$ es una variable aleatoria de dimensión k con vector de medias μ y matriz de covarianzas V semidefinida positiva. Queremos construir una variable aleatoria unidimensional $Y_1 = a_1X_1 + ... + a_kX_k$ con $a_1^2 + ... + a_k^2 = 1$. También queremos que la varianza Var(Y_1)=Var(a'X) sea máxima[2]. Geométricamente, hacemos un cambio de variable para que la dispersión sea máxima y la normalización equivale a mantener la escala original (proyectar). Una solución de este problema será la primera componente principal.

Es decir, la primera componente principal $Y_1 = a'X$ será solución del problema:

$$\left.\begin{array}{ll} \max & Var(a'X) \\ \text{s. a.} & a'a = 1 \end{array}\right\}$$

Dada Y_1, la segunda componente principal no debe contener información ya incluída en Y_1. Por tanto, llamaremos segunda componente principal a una solución de:

$$\left.\begin{array}{ll} \max & Var(a'X) \\ \text{s. a.} & a'a = 1 \\ & Cov(a'X, Y_1) = 0 \end{array}\right\}$$

[2]Nótese que si no normalizamos, $a'a = 1$, la variable Y_1 puede tener una varianza tan grande como queramos.

Por inducción, dadas Y_1, Y_2,...,Y_{k-1} componentes principales, se define la componente principal k-ésima como una solución al problema:

$$\left.\begin{array}{ll} \max & Var(a'X) \\ \text{s. a.} & a'a = 1 \\ & Cov(a'X, Y_j) = 0,\ j = 1, \ldots, k-1. \end{array}\right\}$$

1.4. Teorema de las componentes principales

El teorema siguiente prueba la existencia de las componentes principales y muestra cómo se pueden calcular.

Teorema 1.4.1. *Sea X=($X_1, ..., X_k$) una variable aleatoria k-dimensional con V=Cov(X), matriz de covarianzas, definida positiva. Entonces las (unas) componentes principales son*

$$Y = \begin{pmatrix} Y_1 \\ \ldots \\ Y_k \end{pmatrix} = T'X = \begin{pmatrix} t_{1,1} & \ldots & t_{k,1} \\ \ldots & \ldots & \ldots \\ t_{1,k} & \ldots & t_{k,k} \end{pmatrix} \begin{pmatrix} X_1 \\ \ldots \\ X_k \end{pmatrix}$$

donde T es una matriz ortogonal, $T'T = TT' = I$, tal que T'VT = D = diag(λ_1, λ_2, ..., λ_k) con $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$.

Demostración.

Como V es una matriz simétrica y definida positiva, existe una matriz T=($t_{i,j}$) ortogonal (T'T=TT'=1) tal que T'VT = D =diag(λ_1, λ_2, ..., λ_k) con los valores propios verificando $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$.
De esta forma, si

$$Y = \begin{pmatrix} Y_1 \\ \ldots \\ Y_k \end{pmatrix} = T'X = \begin{pmatrix} t_{1,1} & \ldots & t_{k,1} \\ \ldots & \ldots & \ldots \\ t_{1,k} & \ldots & t_{k,k} \end{pmatrix} \begin{pmatrix} X_1 \\ \ldots \\ X_k \end{pmatrix}$$

entonces Y_1, Y_2, ..., Y_k verifican

$$Cov(Y) = Cov(T'X) = T'Cov(X)T = T'VT = D$$

lo que implica que Cov(Y_i,Y_j)=0 para i ≠ j y Var(Y_j)=λ_j.
Los vectores columnas de T (filas de T') t_j=($t_{1,j}$,...,$t_{k,j}$)' serán una base ortonormal de vectores propios de V verificándose Y_j=t'_jX y Vt_j=$\lambda_j t_j$, para $j = 1, ..., k$.
Veamos que $Y_1 = t_1X$ es una primera componente principal, es decir, que Y_1 es una solución de

$$\left.\begin{array}{ll} \max & Var(a'X) \\ \text{s. a.} & a'a = 1 \end{array}\right\}$$

Supongamos que $a'X$ es una combinación lineal con $a'a = 1$. Entonces, como los vectores propios son una base, existirán $c_1, \ldots, c_k$ números reales tales que

$$a = c_1t_1 + \ldots + c_kt_k,\ con\ c = (c_1, \ldots, c_k),$$

con lo que

$$\begin{aligned} Var(a'X) &= E(a'(X-\mu)(X-\mu)'a) \\ &= a'Va \\ &= \left(\sum_{i=1}^{k} c_i t_i'\right) V \left(\sum_{i=1}^{k} c_i t_i\right) \\ &= \left(\sum_{i=1}^{k} c_i t_i'\right) \left(\sum_{i=1}^{k} c_i V t_i\right) \\ &= \left(\sum_{i=1}^{k} c_i t_i'\right) \left(\sum_{j=1}^{k} c_j \lambda_j t_j\right) \\ &= \sum_{i,j} c_i c_j \lambda j t_i' t_j \\ &= \sum_{i=1}^{k} c_i^2 \lambda_i \end{aligned}$$

y, como

$$a'a = \left(\sum_{i=1}^{k} c_i t_i'\right) \left(\sum_{j=1}^{k} c_j t_j'\right) = \sum_{i,j} c_i c_j t_i' t_j = \sum_{i=1}^{k} c_i^2 = c'c = 1$$

la varianza será máxima si $c_1^2 = 1, c_2 = 0, \ldots, c_k = 0$ ya que

$$Var(t_1'X) = \lambda_1 = \sum_{i=1}^{k} c_i^2 \lambda_1 \geq \sum_{i=1}^{k} c_i^2 \lambda_i = Var(a'X), \forall a \;\; t.q. \;\; a'a = 1.$$

Es decir, $Y_1 = t_1'X$ es una primera componente principal (puede haber otras soluciones si $\lambda_1 = \lambda_2$).
Ahora, por inducción, supongamos que $Y_1 = t_1'X, \ldots, Y_{m-1} = t_{m-1}'X$ son unas (m-1) primeras componentes principales. Veamos que $Y_m = t_m'X$ es la (una) m-ésima componente principal, es decir, es una solución de

$$\left.\begin{array}{ll} \max & Var(a'X) \\ \text{s. a.} & a'a = 1 \\ & Cov(a'X, Y_j) = 0, \; j = 1, \ldots, m-1. \end{array}\right\} \tag{1.5}$$

Como se debe verificar

$$\begin{aligned} Cov(a'X, Y_j) &= Cov(a'X, t_j'X) = E(a'(X-\mu)(X-\mu)'t_j) \\ &= a'Vt_j = \lambda_j a't_j \\ &= \lambda_j \left(\sum_{i=1}^{k} c_i t_i'\right) t_j \\ &= \sum_{i=1}^{k} c_i \lambda_j t_i' t_j \\ &= \lambda_j c_j = 0 \qquad \forall j = 1, \ldots, m-1 \end{aligned}$$

Como dijimos al principio, al ser V definida positiva, $\lambda_1 \geq \ldots \geq \lambda_k > 0$ y entonces de la igualdad anterior llegamos a que necesariamente $c_j = 0 \quad \forall j = 1 \ldots m-1$. $(*_1)$. Como consecuencia se tiene que $\sum_{i=m}^{k} c_i^2 = 1$.$(*_2)$

La varianza será máxima si $c_m = 1$ y $c_j = 0$ para $j > m$, ya que

$$Var(a'X) = \sum_{i=1}^{k} c_i^2 \lambda_i \overset{(*_1)}{=} \sum_{i=m}^{k} c_i^2 \lambda_i \leq \left(\sum_{i=1}^{m} c_i^2 \right) \lambda_m \overset{(*_2)}{=} \lambda_m = Var(\pm t_m X)$$

para todo a tal que $a'a = 1$ y $Cov(a'X, Y_j) = 0$, $j = 1, \ldots, m-1$, es decir, $Y_m = \pm t_m X$ es una solución de (1.1) y por tanto es una m-ésima componente principal.

□

Como consecuencia de la demostración del teorema anterior se obtiene el siguiente corolario:

Corolario 1.4.1. *Sea $\lambda_1 > \lambda_2 > \ldots > \lambda_k$. Entonces las componentes principales son únicas salvo signo.*

Demostración.

El resultado es inmediato, pues si $\lambda_1 > \lambda_2 > \ldots > \lambda_k$, los autovectores serán únicos salvo el signo, pues de norma uno puede haber dos autovectores.

□

Observación 1.4.1. *Nótese que la componente j-ésima se obtiene como $Y_j = t_j' X$, donde $t_j' = (t_{1,j}, \ldots, t_{k,j})$ es un vector propio unitario correspondiente al j-ésimo valor propio (vectores columna de T). Además, $Var(Y_j) = \lambda_j$ y*

$$traza(V) = \sum_{j=1}^{k} \sigma_{j,j} = \sum_{j=1}^{k} Var(X_j) = \sum_{j=1}^{k} Var(Y_j) = \sum_{j=1}^{k} \lambda_j$$

*(las matrices semejantes tienen las trazas iguales), es decir, la variabilidad (información) de las variables originales es la suma de las variabilidades de las componentes principales. La **cantidad de información** (%) contenida en cada componente se calculará como $I_j = 100\lambda_j / (\sum_{i=1}^{k} \lambda_i)\,\%$.*

Por tanto, la traza es una medida unidimensional de la dispersión de una variable k-dimensional. Otra medida es el determinante de la matriz de covarianzas V, $|V| = \lambda_1 \ldots \lambda_k$, que calcula la variabilidad como el área encerrada en el paralelogramo de lados iguales a los valores propios.

Análisis con la matriz de correlaciones

Como hemos visto anteriormente, la forma de obtener las componentes principales es maximizando la varianza de la proyección. Sin embargo, cuando las variables tienen unidades distintas no es conveniente hacerlo así. Si disminuimos la escala de medida de una variable cualquiera, de manera que aumenten en magnitud sus valores numéricos (por ejemplo, pasar de medir en kilómetros a metros), el peso de esa variable en el análisis aumentará.

En resumen, cuando las escalas de medidas de las variables son muy distintas, la maximización de la varianza dependerá decisivamente de estas escalas de medida y las variables con valores más grandes tendrán más peso en el análisis.

Cuando estamos en un caso así, las componentes principales suelen calcularse a partir de la matriz de correlaciones $\prod = (\rho_{i,j}) = \sigma_{i,j}/\sigma_i\sigma_j$, lo que equivale a considerar desde el principio las variables estandarizadas

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

donde se igualan las varianzas a 1.

De esta forma, aplicando el teorema principal sobre $Z_1, \ldots, Z_k$, se obtienen las componentes

$$\widetilde{Y} = \widetilde{T}'Z = \widetilde{T}'diag(V)^{-1/2}(X - \mu)$$

$$\widetilde{Y_j} = \widetilde{t_j}'Z = \sum_{i=1}^{k} \widetilde{t}_{i,j} Z_i = \sum_{i=1}^{k} \widetilde{t}_{i,j} \frac{X_i - \mu_i}{\sigma_i}$$

donde $\widetilde{T}$ es la matriz ortogonal que diagonaliza $\prod = Corr(X) = Cov(Z)$

$$\widetilde{T}'\Pi\widetilde{T} = diag(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k) = \widetilde{D}$$

$\Pi\widetilde{t_j} = \lambda_j\widetilde{t_j}$ y $Z = (Z_1, \ldots, Z_k)'$. De esta forma, se obtiene

$$Cov(\widetilde{Y}) = Cov(\widetilde{T}'Z) = \widetilde{T}'\Pi\widetilde{T} = \widetilde{D}.$$

Es decir, las componentes principales obtenidas a partir de la matriz de correlaciones serán las variables incorreladas con varianza máxima que se pueden obtener a partir de combinaciones lineales de las variables estandarizadas $Z = diag(V)^{-1/2}(X - \mu)$. Los resultados que se obtienen son (en general) diferentes de los que se obtienen a partir de V.

1.5. Propiedades

En la proposición siguiente estudiamos las relaciones entre las nuevas variables y las originales.

Proposición 1.5.1. *Si Y son las componentes principales obtenidas a partir de X, entonces*

$$Cov(X, Y) = TD$$

$$Corr(X, Y) = diag(V)^{-1/2}TD^{1/2}$$

donde diag(V)=diag(σ_1^2,...,σ_k^2).

Demostración.

Por una parte, como

$$Corr(X_i, Y_j) = \frac{Cov(X_i, Y_j)}{\sigma_i\lambda_j^{1/2}}$$

entonces $Corr(X,Y) = diag(V)^{-1/2}Cov(X,Y)D^{-1/2}$. Por otro lado se tiene que

$$Cov(X,Y) = Cov(X,T'X) = VT$$

y, como $T'VT = D$ y T es ortogonal, entonces multiplicando por T a ambos lados de la expresión tenemos que $TT'VT = TD$, por tanto $VT = TD$.

Entonces $Cov(X,Y) = TD$ y $Corr(X,Y) = diag(V)^{-1/2}TD^{1/2}$.

□

Definición 1.5.1. *LLamaremos **matriz de saturaciones** a A=Corr(X,Y).*

Proposición 1.5.2. *Si $\widetilde{Y}$ son las componentes principales obtenidas a partir de la matriz de correlaciones de X, entonces*

$$Corr(X,\widetilde{Y}) = \widetilde{T}\widetilde{D}^{1/2}$$

Demostración.

En efecto, si $\widetilde{Y} = \widetilde{T}'Z = \widetilde{T}'diag(V)^{-1/2}(X-\mu)$, entonces:

$$Cov(Z,\widetilde{Y}) = Cov(Z,\widetilde{T}'Z) = \Pi\widetilde{T} = \widetilde{T}\widetilde{D}$$

$$Corr(X,\widetilde{Y}) = Corr(Z,\widetilde{Y}) = Cov(Z,\widetilde{Y})\widetilde{D}^{-1/2} = \widetilde{T}\widetilde{D}\widetilde{D}^{-1/2} = \widetilde{T}\widetilde{D}^{1/2}$$

□

1.6. Análisis de las componentes principales

Veamos algunos tópicos relevantes para realizar un PCA:

Componentes de tamaño y forma

Cuando existe una alta correlación y positiva entre todas las variables, la primera componente principal tiene todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables, o un factor global de "tamaño". Las restantes componentes se interpretan como factores "de forma" y típicamente tienen coordenadas positivas y negativas, que implica que contraponen unos grupos de variables frente a otros. Estos factores de forma pueden frecuentemente escribirse como medias ponderadas de dos grupos de variables con distinto signo y contraponen las variables de un signo a las del otro.

La interpretación de las componentes se simplifica suponiendo que los coeficientes pequeños son cero y redondeando los coeficientes grandes para expresar la componente como cocientes, diferencias o sumas entre variables. Estas aproximaciones son razonables si modifican poco la estructura de la componente y mejoran su interpretación.

Selección de números de componentes

Una vez que hemos realizado el PCA surge la pregunta sobre cuántas componentes principales son las que debemos utilizar. La respuesta no es única, pues depende de factores subjetivos.

Todas las soluciones son correctas, pues lo que estamos haciendo es perder información (la menor posible) a cambio de reducir la dimensión inicial. Existen diferentes técnicas para seleccionar el número de componentes. Sin embargo, en vista de que el objetivo es crear PCA-Gapminder[3], tomaremos siempre las dos primeras. Será fundamental conocer la información total mantenida. Si se advierte que el número es bajo se considerará añadir la tercera y la cuarta componentes al estudio.

Representación gráfica

La interpretación de las componentes principales se favorece representando las proyecciones de las observaciones sobre un espacio de dimensión dos, definido por parejas de las componentes principales más importantes. La proyección de cualquier observación sobre una componente es directamente el valor de la componente para esa observación. La representación habitual es tomar dos ejes ortogonales que representen las dos componentes consideradas, y situar cada punto sobre ese plano por sus coordenadas con relación a estos ejes, que son los valores de las dos componentes para esa observación.

Análisis de las componentes principales en RStudio

Una vez cargado en RStudio el conjunto de *datos* sobre el que se quiere realizar el PCA, es conveniente hacer un primer estudio de las variables por separado para detectar datos atípicos, falta de simetría o normalidad. Para ello, usamos el comando *summary(datos)*. Esto nos muestra las medias, medianas, cuartiles, mínimos y máximos.

Por otra parte, RStudio también ofrece la opción de estudiar las relaciones entre las variables mediante *boxplot(datos)*. Los gráficos de caja-bigote muestran si existe alguna variable que tenga mucha más dispersión que las demás. Podemos calcular las matrices de correlación y covarianza con *cov(datos)* y *cor(datos)*. Con ayuda de estos métodos decidiremos si se observa que las variables se miden en unidades bastante diferentes. Si es así, utilizaremos la matriz de correlación para hacer el PCA (estandarizando las variables).

Asumiremos en este proyecto el uso de la matriz de correlaciones para el cálculo del PCA, ya que en el capítulo tercero los datos de los casos prácticos se miden en unidades muy distintas y estaremos en este caso. Para calcular las componentes principales usaremos el comando *PCA<-princomp(datos, cor=TRUE)*.

Para ver las características principales haremos *summary(PCA, loadings=TRUE)*. Obtendremos tres filas de valores para cada componente: la desviación estándar (raíces cuadradas de de los valores propios de la matriz de correlaciones ordenados de mayor a menor), la proporción de sus varianzas y las proporciones acumuladas. Todas ellas miden la importancia de las componentes. Las proporciones acumuladas se calculan sumando las de las componentes anteriores. Estos valores nos indican en tanto por uno la información mantenida respecto de la información inicial.

Las cargas o *loadings* son los vectores propios unitarios de los valores propios. Podemos calcularlas usando el comando *T<-PCA$loadings*. Tecleando *T[,i]*, siendo i el número de la componente que queremos estudiar, obtendremos el vector propio utilizado para calcularla. De esta forma, las componentes principales tendrán la forma:

[3]Utiliza gráficos bidimensionales. Se verá en el capítulo posterior.

$$Y_j = a_1 * X_1^* + a_2 * X_2^* + \ldots + a_p * X_p^*$$

con los a_i siendo las cargas (coeficientes) y las variables estandarizadas $X_i^* = \frac{X_i - mean(X_i)}{sd(X_i}$.

Por otra parte, las puntuaciones o *scores* son los valores obtenidos para los individuos de la muestra en las componentes principales. Mediante el comando *S<-PCA$scores* las conseguiremos en RStudio.

Si representamos las cargas y las puntuaciones de dos componentes (normalmente las dos primeras, ya que son las que más información contienen), obtendremos las cargas vistas como vectores. Este gráfico será la mejor proyección visual o "foto" de los ejes iniciales de todas las variables estandarizadas y de las puntuaciones de los individuos de la muestra.

Las variables con vectores largos (norma cercana a 1) estarán bien representadas por esas dos componentes principales, mientras que las que tengan vectores cortos estarán mal representadas (se pierden al proyectar por ser casi perpendiculares). Las saturaciones se usarán para saber cuánta información se mantiene (o se pierde) de cada variable en cada componente. Las puntuaciones se usarán para decir cómo serán los individuos de la muestra en esas características.

En el capítulo tercero utilizaremos el comando *biplot* en RStudio para obtener esa "foto" y analizaremos las componentes principales en función de cómo sean los vectores. A partir del análisis de estos vectores podremos determinar qué variables están mejor representadas y dar así un significado aproximado de cada componente principal.

Capítulo 2

Marco práctico. PCA-Gapminder

2.1. Hans Rosling: el destructor de mitos

"Desmontando mitos sobre el mundo": así tituló Eduardo Punset el programa[1] que dedicó a Hans Rosling cuyo objetivo literalmente era "revelar la cara fascinante de los números y las estadísticas, y su inmenso poder para explicar el pasado y el futuro del mundo"

Hans Rosling[2] fue un médico sueco que, muy joven, se trasladó a Mozambique donde llegó a ser el único facultativo de una población de 300.000 personas. Contribuyó a la identificación y prevención del konzo, una enfermedad que dejaba paralíticos a decenas de miles de niños en zonas pobres de África. Pero su visión era global, identificando finalmente que la mayor causa de enfermedad era la pobreza, la ignorancia y la injusticia. La relación entre salud y pobreza fue una de sus obsesiones desde entonces. Sólo la pérdida de un hijo le hizo volver con toda su familia a Suecia, donde sería profesor de salud pública para los muy afortunados alumnos de Uppsala y luego de Estocolmo. Fue, eso sí, un hombre positivo, optimista e ilusionado por lo que se puede hacer en el mundo en la lucha contra la enfermedad y la pobreza.

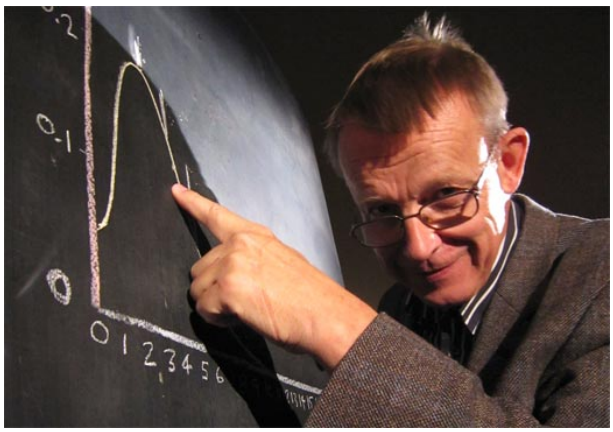


Figura 2.1: Hans Rosling.

En este punto conviene que expliquemos la razón por la que un médico aparece en la introducción de un proyecto fin de grado de la Facultad de Matemáticas.

El objetivo de Rosling fue difundir los datos sobre nuestro mundo de una forma moderna, muy visual, muy gráfica, muy espectacular, en movimiento, en tres dimensiones; original y entusiasta, pero sin perder rigor técnico. Fue un gigante de la divulgación.

Le sorprendía cómo la mayor parte de las personas tienen unas ideas erróneas sobre las relaciones internacionales, el desarrollo, la pobreza y la salud. Y se propuso mejorar esa circuns-

[1] Programa "Redes" emitido el 30 de mayo de 2010, disponible en rtve.es.
[2] Biografía obtenida del artículo referido en la fuente [2] citada en la Bibliografía.

tancia. Fue consejero de la OMS, UNICEF y otras agencias de ayuda humanitaria; iniciador de Médicos Sin Fronteras en Suecia... Rosling puso en marcha la Fundación Gapminder que desarrolló Trendalyzer, un programa para convertir series estadísticas en gráficos interactivos; para dar movimiento a los cientos de miles de cifras enterradas en bases de datos colosales, como las del Banco Mundial. Su pretensión era evidente: promover una visión del mundo basada en hechos y datos, a través de la comprensión de información estadística pública.

Hay vidas que cambian en solo un instante. Eso fue lo que le ocurrió a Hans Rosling. Fue en una charla TED[3] en 2006 cuando dio a conocer al público su talento como conferenciante con ayuda de su innovadora herramienta gráfica Gapminder. En 2007, Google adquirió Trendalyzer.

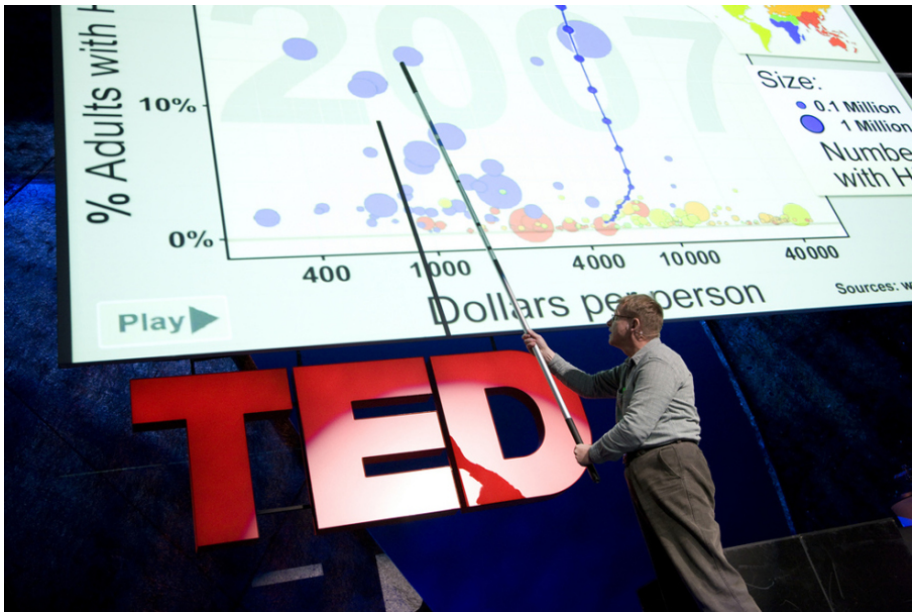


Figura 2.2: Hans Rosling en una de sus charlas TED.

2.2. Origen de Gapminder

Gapminder es una fundación que fue creada el 25 de febrero de 2005 en Estocolmo por una familia sueca. Los miembros de esta familia que participaron en el proyecto fueron Hans Rosling, su hijo Ola Rosling y su nuera Anna Rosling.

Es una fundación sin fines de lucro. Afirman que tratan de dar una visión del mundo basada en los hechos (for a fact-based world view) poniendo a nuestra disposición un método que nos haga más fácil entender los datos.

Para ello nos ofrecen una herramienta gráfica que permite analizar diferentes tipos de datos numéricos de distintos países en forma de secuencia temporal. Los datos que podemos encontrar dentro de Gapminder van desde el número de casos de VIH, la esperanza de vida, la malnutrición infantil, el número de hijos que ha tenido una mujer, las emisiones de CO2, la edad del primer

[3]TED: "Tecnología, Entretenimiento, Diseño" es una organización sin fines de lucro dedicada a ideas dignas de difundir (del inglés: ideas worth spreading). TED es ampliamente conocida por sus charlas que cubren un amplio espectro de temas que incluyen ciencias, arte y diseño, política, educación, cultura, negocios, asuntos globales, tecnología, desarrollo y entretenimiento. - Fuente: Véase [8] en Bibliografía.

matrimonio, renta per capita, producto interior bruto, gasto energético, consumo energético, ... En total, 503 indicadores que podemos analizar.

La herramienta nos permite correlacionar de forma sencilla cualquiera de los indicadores y mediante una animación tener la posibilidad de analizar la evolución temporal. Además, podemos elegir los países que queremos estudiar.

En una entrevista a Hans Rosling publicada en El País aseguraba que *"la idea de filtrar datos, transformarlos en información para que un público amplio entienda hechos complejos, es el objetivo de Gapminder. Si el usuario medio ve solo los datos en bruto, no entenderá nada. Pero si accede a ellos después de procesados, el efecto es muy diferente. Pueden cambiar su forma de pensar sobre muchos asuntos."*

El software estadístico Gapminder World es un visualizador interactivo de gráficos y mapas que permite observar la situación de múltiples individuos mediante una extensa gama de variables. Gapminder World presenta un gráfico bidimensional mediante el que los individuos se posicionan en base a las dos variables seleccionadas, como burbujas en diferentes colores y tamaños, según las características de estos. Además este software estadístico introduce la variable tiempo para observar años concretos o para obtener imágenes dinámicas de la evolución experimentada por los individuos en un gráfico animado.

2.3. PCA-Gapminder en RStudio

Percatada la utilidad de Gapminder, surgió la idea de complementarla con el análisis de componentes principales. Con ayuda de ambas se podría hacer un estudio multivariante muy completo. Es por esto que le damos el nombre de PCA-Gapminder.

2.3.1. Características de la muestra

En primer lugar, describiremos cómo debe ser la muestra ideal para exprimir al máximo las posibilidades que ofrece el programa. Lo más habitual es que sea una tabla, donde aparecen los valores de p variables observadas sobre n individuos. Sin embargo, el efecto dinámico que tiene Gapminder obliga a recoger de cada individuo las observaciones de las p variables en los r años que comprende el periodo sobre el que se realiza el estudio. Por tanto, incluiremos a esa estructura habitual una variable nueva llamada *"Año"* (o mes, semestre, etc.), de tal manera que se tendrán $p \cdot r$ datos recogidos como individuos y $p+1$ variables.

Debe constar de variables de tipo cuantitativo (su valor se expresa numéricamente) y cualitativo (su valor es un atributo o una categoría). Las primeras son las que se usarán para el cálculo de la parte más analítica, el análisis de componentes principales. Las segundas, sin embargo, servirán como elementos descriptivos en la representación gráfica. Más adelante veremos que serán representadas con el color y el tamaño de las burbujas. Por otro lado, la muestra siempre debe incluir una variable temporal, la parte dinámica de PCA-Gapminder vendrá determinada por ella.

2.3.2. Extensión de Gapminder a PCA-Gapminder

PCA-Gapminder es un gráfico bidimensional dinámico que representa a los individuos de un estudio a través de burbujas que se van moviendo por la pantalla a medida que el tiempo

va a avanzando. Se pueden escoger las variables que se quieren estudiar. Además incorpora dos técnicas muy visuales, el color y el tamaño de las burbujas que, a su vez, representarán otras dos variables, siendo el color una variable identificativa de un grupo y el tamaño otra variable de interés.

Hasta ahora todos estos detalles eran los que ofrecía Gapminder. La novedad viene a la hora de elegir qué se representa en los ejes del gráfico bidimensional. Gapminder daba la opción de elegir cualquier variable y enfrentarla a otra variable del estudio, con lo que el resto de variables quedaban sin representar. Sin embargo, en nuestro trabajo nos movemos en el ámbito multivariante, donde la cantidad de variables es considerablemente grande. No es práctico tener que ir cambiando las variables de eje para ver cómo se relacionan unas con otras o para obtener una visión más global. Estas mejoras son las que propone PCA-Gapminder con ayuda del análisis de componentes principales. En nuestra nueva herramienta se plantea la idea de representar en los ejes las componentes principales. Por tanto, las variables cuantitativas de la muestra, serán las que intervengan en el cálculo de las componentes principales. Una vez calculadas Y_1 e Y_2, y habiéndolas interpretado, podremos representarlas en PCA-Gapminder y extraer conclusiones de manera más global.

En la Sección 1.6 del Capítulo primero ya se comentó que a pesar de que existen diversas técnicas para decidir el número de componentes principales seleccionadas, para nuestro programa necesitaremos pares de componentes. En principio usaremos las dos primeras para el gráfico bidimensional, pues son las que más información poseen. Sin embargo, si observamos que la proporción acumulada es baja, recurriremos a incrementar el número de componentes principales seleccionadas (Y_3 e Y_4 probablemente) y combinaremos unas con otras para estudiar los resultados obtenidos.

Podríamos preguntarnos para qué hace falta incorporar esta nueva técnica, cuando ya contamos en RStudio con una representación bidimensional de dos componentes principales dada por el comando *biplot*. Más adelante, en el capítulo tres, veremos sobre ejemplos concretos que esa representación dada se apelotona al mostrar juntos todos los datos y no aporta información clara. Tampoco muestra las variables cualitativas de las que hemos hablado anteriormente aportada por Gapminder ni, por supuesto, el dinamismo temporal.

Antes de poder interpretar el significado de la ejecución, se tendrá que obtener el significado o una explicación aproximada de lo que es cada componente principal.

2.3.3. Programación de PCA-Gapminder

R es un potente lenguaje orientado a objetos y destinado al análisis estadístico y la representación de datos. Se trata de un software libre que permite su utilización gratuitamente. La comunidad científica internacional lo ha elegido como la "lingua franca" del análisis de datos, y tiene una gran implantación en universidades y cada vez más en el mundo empresarial. La programación de PCA-Gapminder se va a desarrollar con RStudio, un entorno desarrollado integrado para R.

Con ayuda de RStudio, es posible implementar nuestras propias gráficas dinámicas, emulando así las técnicas Gapminder de Hans Rosling. Para ello, se puede recurrir a aplicaciones de Google (Google apps), para generar código html que contenga la gráfica deseada.

En primer lugar, hay que cargar en el script el paquete *googleVis*, que permite usar las gráficas de Google apps. Dentro de todas las opciones que ofrece, los gráficos más conocidos son precisamente los *Motion Chart*, popularizados por Hans Rosling en sus charlas TED. Las funciones de este paquete permiten al usuario visualizar datos almacenados en RStudio con *Google Charts* sin tener que subir los datos a Google. La salida de una función de *googleVis* es código html que contiene datos y referencias a funciones JavaScript recibidas por Google. Para poder reproducir la ejecución es necesario un navegador con conexión a internet.

Cuando ya se ha cargado dicho paquete, *library(googleVis)*, hay que introducir los datos procedentes de un fichero Excel con los que vamos a hacer el estudio. Lo haremos mediante el comando *read.table*, apropiado para leer este tipo de archivos .xls o .csv. Dentro de la función especificaremos el nombre del archivo; indicaremos la presencia de números decimales para que interprete la coma como tal y no como un carácter de código alfabético; y finalmente, señalaremos que las columnas tienen un encabezado con el nombre de la variable. Todo ello se expresará en RStudio como sigue.

read.table(file=" archivo.csv", dec=",", header=TRUE).

Ya tenemos preparados los datos y la librería necesaria. A continuación comienza la programación de la parte analítica, el cálculo de las componentes principales. En Gapminder se representaban las variables en los ejes; sin embargo, el nuevo objetivo es hacerlo con las componentes principales. Lo que representaremos en los ejes serán las puntuaciones de cada individuo. Para ello necesitamos primero calcular las componentes. Una vez hecho esto, actualizaremos el contenido del objeto de datos con el que estamos trabajando, creando nuevas columnas que contengan el valor de las puntuaciones de cada individuo de la muestra para cada componente principal de las que hayamos seleccionado.

En la Sección 1.6 del Capítulo primero se señalaron los comandos más importantes para hacer el PCA en RStudio, y también que utilizaríamos la matriz de correlaciones por el cambio de escala en las componentes. Se dispone del comando *princomp* para el cálculo de las componentes principales, al cual hay que pasarle el objeto de datos que se leyó del fichero Excel y una indicación de que vamos a usar la matriz de correlaciones.

princomp(datos, cor=TRUE)

Guardamos esto en un objeto al que llamamos *PCA*. Ahora es preciso encontrar una manera aproximada de definir a las componentes principales para posteriormente poder interpretar el gráfico obtenido. Para ello estudiaremos las cargas con *PCA$loadings*.

Por otro lado, hay que almacenar las puntuaciones de cada componente principal en un vector. Guardamos todas las puntuaciones primeramente en un objeto llamado *S*, con el comando *PCA$scores*, y como en principio se usarán sólo las dos primeras componentes principales, haremos

S[,1]-> Y1
S[,2]-> Y2

para así guardar estos nuevos vectores columna como nuevos objetos y a continuación poder pegarlos en el fichero de datos inicial. Esto lo haremos escribiendo

cbind(archivo, Y1, Y2)->FicheroNuevo

Finalmente, llamaremos a la función que creará un objeto, en código html, que es el gráfico de burbujas buscado. Hay que indicarle que tiene que utilizar el nuevo fichero de datos en el que hemos añadido las puntuaciones, quién es la variable identidad que van a representar las burbujas y quién es la variable temporal.

gvisMotionChart(FicheroNuevo,idvar="individuos",timevar="Año")->grafico
plot(grafico)

Le damos nombre al objeto (*grafico*) para finalmente representarlo mediante el comando *plot(grafico)*. Al ejecutar la línea en RStudio se abrirá automáticamente una pestaña en el navegador con el gráfico dinámico buscado. En la siguiente sección explicaremos las opciones que ofrece el programa una vez ejecutado.

2.3.4. Aclaraciones sobre la información de pantalla

La Figura 2.3 que se muestra a continuación es la pantalla que obtenemos del programa PCA-Gapminder cuando se haya ejecutado. A continuación se van a comentar los detalles que ofrece la aplicación y el significado de todos los elementos que aparecen.

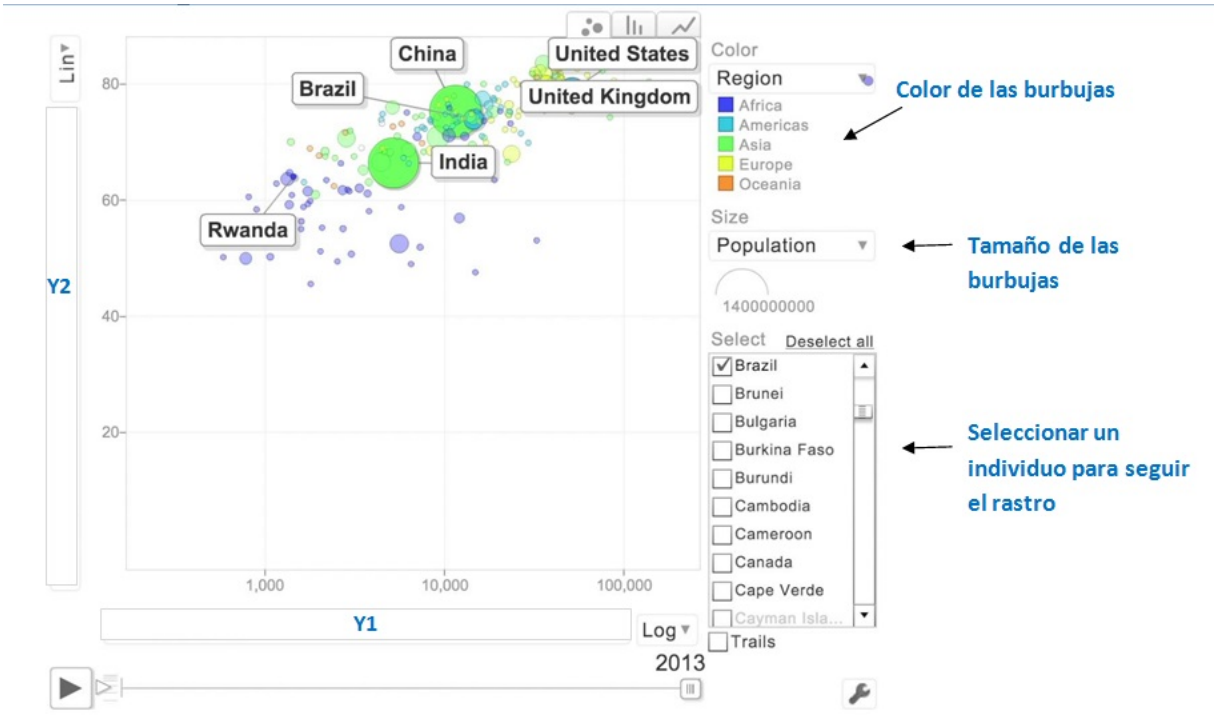


Figura 2.3: Pantalla de PCA-Gapminder.

Las burbujas representan los individuos de la muestra. En este ejemplo ilustrativo son los países del mundo. Hay dos variables descriptivas, el color y el tamaño de dichas burbujas. Con el color se suele simbolizar una variable que agrupe a los individuos por sectores, en este caso el color los organiza según el continente al que pertenecen. Respecto al tamaño, normalmente

se usa para la población. Se observa que las burbujas grandes representan países como India, China o Estados Unidos.

La línea horizontal que está debajo del gráfico es la recta temporal. Al pulsarle al botón "play"(triángulo a la izquierda de la línea) el dinamismo del gráfico comenzará, recorriendo el periodo de tiempo establecido.

En los ejes se representan las componentes principales Y1 e Y2. En los casos necesarios también podremos representar Y3 e Y4. Previamente a la representación habremos hecho un análisis para deducir el significado de las componentes. Así, se estará en condiciones de interpretar el gráfico.

En la Figura 2.3 se ve también una flecha que indica dónde se puede seleccionar un individuo en concreto para seguirle el rastro. Esta opción será interesante cuando advirtamos que alguna burbuja tiene un comportamiento muy diferente de las demás. Podremos rastrearla y estudiarla ese caso en específico.

Capítulo 3

Análisis de datos con PCA-Gapminder

3.1. Caso 1. Asignaturas en el grado en matemáticas

Para hacer el estudio de este caso, en primer lugar, se recogió una muestra de datos[1] sobre las asignaturas del primer y segundo curso del Grado en Matemáticas de la Universidad de Murcia. Una vez que se obtuvo esta información, el siguiente paso fue la adaptación de estos datos a un modelo que se adecuase a las necesidades que tendría nuestro PCA-Gapminder.

Cada asignatura tenía que tener una variable temporal (año en el que había sido recogida la muestra de variables) y algunas descriptivas (rama y curso), visuales en Gapminder y, por supuesto, las numéricas, aquellas que utilizaríamos posteriormente para realizar el PCA. El formato de la muestra diseñada quedó finalmente así:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Asig.	**Año**	**Curso**	**Rama**	**Matric.**	**Apr.**	**Susp.**	**Nota media**	**Tasa apr.**
FVRI	2010	1	Análisis	90	48	42	4,71	0,533
FVRI	2011	1	Análisis	102	36	66	3,36	0,352
FVRI	2012	1	Análisis	117	37	80	3,4	0,316
FVRI	2013	1	Análisis	101	26	75	3,34	0,257
FVRI	2014	1	Análisis	115	40	75	4,09	0,347
FVRI	2015	1	Análisis	121	38	83	3,73	0,314
...	...	...	...	...	...	...	...	...
GYA	2010	2	Álgebra	25	9	16	6,75	0,36
GYA	2011	2	Álgebra	40	14	26	6,76	0,35
GYA	2012	2	Álgebra	42	21	21	5,01	0,500
GYA	2013	2	Álgebra	55	45	10	5,98	0,818
GYA	2014	2	Álgebra	52	11	41	3,76	0,211
GYA	2015	2	Álgebra	73	44	29	4,47	0,602

Cuadro 3.1: Plantilla de la muestra utilizada.

[1]Proporcionada por ATICA, Área de Tecnologías de la Información y las Comunicaciones Aplicadas de la Universidad de Murcia, y por la secretaría de la Facultad de Matemáticas.

Demos ahora una breve descripción de cada variable:

- Asignatura: Son los "individuos" sobre los que gira el estudio. Hay 20, 10 correspondientes al primer curso y otras 10 de segundo curso.
- Año: Esta variable toma los valores entre 2010 y 2015. Como hemos decidido que la variable temporal sea anual, los resultados relativos a cada asigntaura se consideraran uniendo las tres convocatorias. Es decir, tomaremos por ejemplo el número de aprobados sumando los que superaron la asignatura en febrero, en junio o julio/septiembre. Por otro lado, los suspensos se cuentan restando a final de curso el número de aprobados al de matriculados.
- Curso: Es una variable cualitativa. Puede tomar los valores 1 o 2 según al curso al que corresponda esa asignatura.
- Rama: Es cualitativa. Será la variable identificativa que mediante el color nos de una visión más clara al hacer PCA-Gapminder. Puede valer: *álgebra*, *análisis*, *estadística*, *investigación operativa*, *topología* y *otras*.
- N° matriculados: Esta variable numérica indica el número de personas matriculadas en esa asignatura ese curso.
- N° aprobados: Es numérica. Incluye la cantidad de personas que superaron esta asignatura durante cualquiera de las convocatorias de ese año.
- N° suspensos: Esta variable incluye el número de personas que no superaron la asignatura tras las tres convocatorias, bien porque suspendieron, o bien porque no se presentaron.
- Nota media: Toma un valor numérico que es la nota media calculada entre las tres convocatorias del curso. Incluye todas las notas, suspensas y aprobadas.
- Tasa aprobados: Esta variable es el cociente entre el número de aprobados y el de matriculados en la asignatura.

A continuación realizaremos el estudio de los datos y el análisis de las componentes principales haciendo uso de la herramienta RStudio.

Estudio inicial de los datos

Comenzamos introduciendo los datos en RStudio mediante el comando *read.table*, apropiado para leer información proviniente de un fichero excel. Llamaremos *data* al objeto que contiene nuestro conjunto de datos. Durante todo el análisis, para realizar el PCA, utilizaremos sólo las variables de tipo numérico que son de la 5 a la 9 (número de matriculados, número de aprobados, número de suspensos, nota media y tasa de aprobados), pues las demás únicamente serán necesarias cuando tratemos la parte de Gapminder. Esto se indica en RStudio como *data[5:9]*. Una vez hecho esto, haremos un primer análisis.

Haciendo *summary(data[5:9])* se obtiene un resumen de las variables, incluyendo para cada una de ellas el primer, segundo y tercer cuartil muestral, el máximo, mínimo y la media.

MATRICUL	APROBADOS	SUSPENSOS	NOTA_MEDIA	TASA_APROB
Min: 25.00	Min: 9.00	Min: 5.00	Min: 3.210	Min: 0.2115
1st. Qu. : 51.00	1st. Qu. : 28.75	1st. Qu. : 17.75	1st. Qu. : 4.420	1st. Qu. :0.3682
Median: 63.50	Median: 34.00	Median: 29.50	Median: 4.940	Median: 0.5153
Mean: 72.23	Mean: 33.77	Mean: 37.73	Mean: 4.913	Mean: 0.5097
3rd. Qu.: 99.00	3rd. Qu.: 39.25	3rd. Qu.: 55.50	3rd. Qu.: 5.395	3rd. Qu.:0.6301
Max: 135.00	Max: 61.00	Max: 106.00	Max: 7.700	Max: 0.8529

Cuadro 3.2: Resumen sobre las variables.

Es conveniente hacer un estudio previo de las correlaciones existentes entre las variables. Estas relaciones se verán reflejadas posteriormente en las componentes principales. Podemos verlo de manera gráfica usando *plot(data[5:9])*.

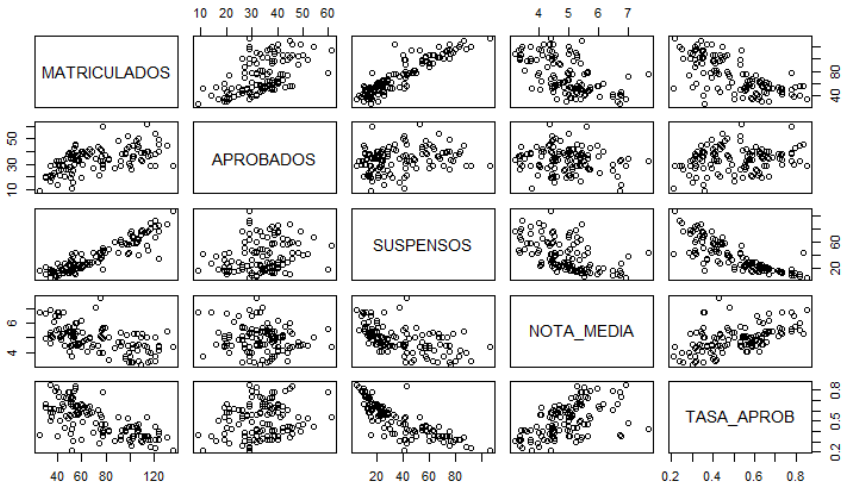


Figura 3.1: Gráficos bidimensionales para las 5 variables.

o bien de forma más analítica, visualizando la matriz de correlaciones, *cor(data[5:9])*

	MATRICUL	APROB	SUSPEN	NOTA_MEDIA	TASA_APROB
MATRICUL	1.0000000	0.5555467	0.9118342	-0.5350756	-0.6744924
APROB	0.5555467	1.0000000	0.2511441	-0.1376782	0.1808540
SUSPEN	0.9118342	0.2511441	1.0000000	-0.5343211	-0.8176592
NOTA_MEDIA	-0.5350756	-0.1376782	-0.5343211	1.0000000	0.5340958
TASA_APROB	-0.6744924	0.1808540	-0.8176592	0.5340958	1.0000000

Cuadro 3.3: Correlaciones entre las cinco variables.

Se observa que existen variables con correlaciones positivas y negativas. Podemos destacar, por ejemplo, la variable número de suspensos frente al número de matriculados ya que tienen una correlación positiva alta. Además de esta matriz podemos estudiar, por otro lado, la matriz de covarianzas mediante el comando *cov(data[5:9])*. Esta matriz es siempre simétrica respecto a su diagonal principal.

	MATRICUL	APROB	SUSPEN	NOTA_MEDIA	TASA_APROB
MATRICUL	811.39048	149.870028	609.55854	-13.5807843	-3.07685665
APROB	149.87003	89.692997	55.81961	-1.1618207	0.27429803
SUSPEN	609.55854	55.819608	550.76863	-11.1733053	-3.07306809
NOTA_MEDIA	-13.58078	-1.161821	-11.17331	0.7939434	0.07621303
TASA_APROB	-3.076857	0.274298	-3.073068	0.07621303	0.02564664

Cuadro 3.4: Matriz de covarianzas de las cinco variables.

El problema es que las escalas (varianzas) son muy diferentes, por eso debemos utilizar la matriz de correlaciones $M < -cor(data[5:9])$ para hacer el PCA.

Cálculo de las componentes principales

Hemos decidido usar la matriz de correlaciones. Con esto las variables originales se estandarizan y todas tienen la misma importancia (varianza 1). El comando que usaremos para hacer el PCA será $PCA < -princomp(data[5:9],cor=TRUE)$. Veamos las características principales. Escribiendo en RStudio *summary (PCA, loadings=TRUE)* se obtiene:

Importance	**Comp. 1**	**Comp. 2**	**Comp. 3**	**Comp. 4**	**Comp. 5**
Standard deviation	1.762339	1.1036977	0.7602414	0.28081502	0.138518082
Proportion of Variance	0.621168	0.2436297	0.1155934	0.01577142	0.003837452
Cumulative Proportion	0.621168	0.8647977	0.9803911	0.99616255	1.000000000

Cuadro 3.5: Importancia de las componentes principales.

La importancia de cada componente principal se mide con las desviaciones estándar. Éstas son la raíces cuadradas de los valores propios de la matriz de correlaciones M ordenados de mayor a menor. Como se explicó anteriormente, al elegir la matriz de correlaciones para hacer nuestro análisis, todas las variables se estandarizaron para tener la misma importancia a priori, es decir, varianza 1. Entonces las varianzas iniciales suman 5 (la traza de la matriz de correlaciones) por lo que los valores de las proporciones de la varianza se calculan de la siguiente manera:

- Componente 1: $0{,}621168 = \frac{1{,}762339^2}{5}$
- Componente 2: $0{,}2436297 = \frac{1{,}1036977^2}{5}$
- Componente 3: $0{,}1155934 = \frac{0{,}7602414^2}{5}$
- Componente 4: $0{,}01577142 = \frac{0{,}28081502^2}{5}$
- Componente 5:$0{,}003837452 = \frac{0{,}138518082^2}{5}$

Las proporciones acumuladas de cada componente se calculan sumando las de las anteriores. Están expresadas en tanto por uno. Por tanto, podemos extraer de esto que la primera componente principal conserva un 62,1168 % de la información inicial, y la primera y la segunda componentes juntas conservan un 86,47977 %.

Las cargas son los vectores propios unitarios de los valores propios anteriores. Estas cargas son los coeficientes de las componentes que queremos calcular.

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
MATRICULADOS	-0.5372760	0.23678485	-0.19977158	-0.06853123	0.781448812
APROBADOS	-0.1924174	0.84714756	0.04680082	-0.28544371	-0.402054868
SUSPENSOS	-0.5399625	-0.04921365	-0.29716834	0.69256680	-0.371565193
NOTA_MEDIA	0.4047701	0.12338719	-0.90397908	-0.06114581	0.004449813
TASA_APROB	0.4678800	0.45675766	0.22894062	0.65607757	0.299347866

Cuadro 3.6: Cargas (loadings) de las componentes principales.

Análisis de las componentes principales

Nuestra decisión acerca de cuáles son las componentes principales con las que vamos a trabajar es clara. Nos quedaremos con las dos primeras, que son las que más información contienen. Ahora bien, necesitamos analizarlas para ver qué información podemos extraer. Mirando las cargas de las componentes que estamos analizando podemos dar significado a éstas. Por tanto, la primera componente principal se calcula como

$$Y_1 = -0{,}5372760X_1^* - 0{,}1924174X_2^* - 0{,}5399625X_3^* + 0{,}4047701X_4^* + 0{,}4678800X_5^*$$

donde $X_i^* = \frac{X_i - mean(X_i)}{sd(X_i)}$ es la variable i-ésima estandarizada.

En primer lugar, recordemos quién era cada variable:

X_1	Alumnos matriculados
X_2	Alumnos aprobados
X_3	Alumnos suspensos/no presentados
X_4	Nota media
X_5	Tasa de aprobados

Cuadro 3.7: Identificación de las variables.

Es interesante para facilitar la interpretación de Y_1 cambiarla de signo.

$$Y_1 = 0{,}5372760X_1^* + 0{,}1924174X_2^* + 0{,}5399625X_3^* - 0{,}4047701X_4^* - 0{,}4678800X_5^*$$

Veamos qué significado tienen estos coeficientes. Los más significativos son los que acompañan a X_1^*, X_3^*, X_4^* y a X_5^*. El primero, *0,5372760* indica que los individuos, las asignaturas, tienen muchos alumnos matriculados. Por otro lado, la carga que acompaña a X_3^*, *0,5399625*, indica que el número de suspensos es alto y, por tanto, la tasa de aprobados será baja (*-0,46788*). Además el coeficiente de X_4^*, *-0,4047701*, indica que las asignaturas se caracterizan por tener notas medias bajas. Por tanto, la componente Y_1 muestra cómo cuando hay muchos matriculados en las asignaturas aumentan los suspensos (disminuyendo la tasa de aprobados como es obvio) y las notas medias son bajas.

Analicemos ahora la segunda componente principal. Utilizando las cargas de éstas que aparecen en el cuadro 3.6 podemos calcularla del siguiente modo:

$$Y_2 = 0{,}23678485X_1^* + 0{,}84714756X_2^* - 0{,}04921365X_3^* + 0{,}12338719X_4^* + 0{,}45675766X_5^*$$

Destacan en este caso las cargas que acompañan a X_2^* y a X_5^*. Son coeficientes altos y positivos. Podemos determinar que esta componente describe asignaturas con un número de aprobados y una tasa de aprobados altos. Interpretaremos esta componente como lo fácil que resulta una asignatura para superarla.

Por tanto, para posteriormente manejar con facilidad estas nuevas variables daremos una "definición" aproximada de lo que representan, en función del estudio de las cargas que hemos hecho anteriormente.

$Y_1 =$ *Asignaturas con muchos matriculados, muchos suspensos y notas bajas.*

$Y_2 =$ *Lo "fácil" que es una asignatura.*

Además de estudiar por separado las componentes principales, también es conveniente ver una representación bidimensional de Y_1 frente a Y_2[2]. Utilizando el comando que sigue a continuación se obtiene el gráfico de la Figura 3.2.

$$biplot(PCA, pc.biplot = TRUE)$$

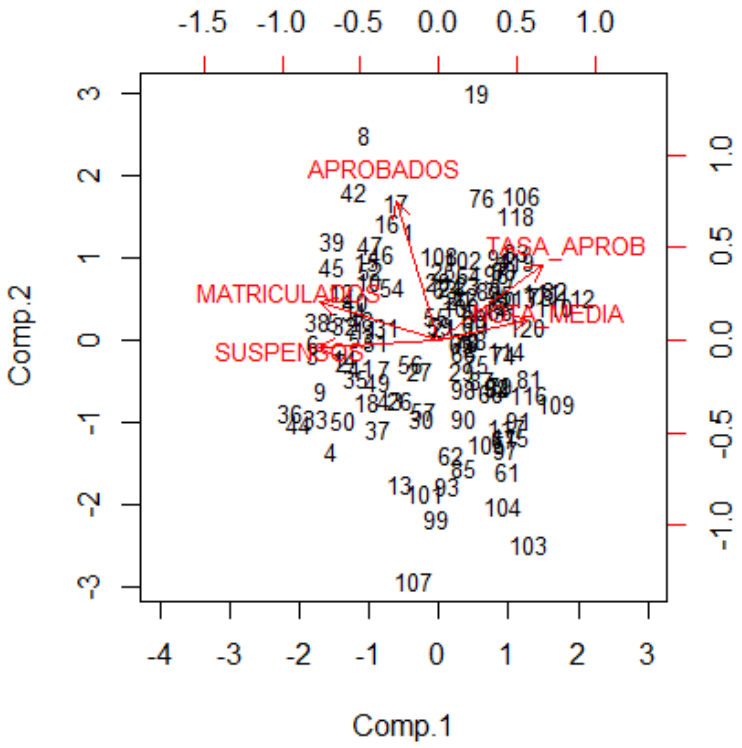


Figura 3.2: Gráfico de las dos primeras componentes.

[2]La representación aquí dada se hace sin hacer el cambio de signo a Y_1.

Los vectores rojos son las cargas, con las escalas en la derecha y arriba. Las puntuaciones están representadas en negro con las etiquetas de los datos (las asignaturas) con las escalas abajo y a la izquierda (Y_1 e Y_2).

Este gráfico es la mejor "foto"de los ejes iniciales de las cinco variables estandarizadas y de las puntuaciones de las asignaturas (individuos de la muestra). Además de analíticamente, como hemos hecho anteriormente, también podemos interpretar las componentes principales mediante este gráfico. Las variables con vectores largos (los de norma cercana a 1) estarán bien representadas en las dos componentes, mientras que las que tengan vectores cortos estarán mal representadas, pues se pierden al proyectar por ser casi perpendiculares.

En nuestro gráfico todas están bien representadas. Más concretamente, observamos que las variables *tasa de aprobados* y la *nota media* hacen crecer la primera componente, mientras que el número de matriculados y el de suspensos la hacen disminuir. La variable de aprobados apenas influye en ella. La interpretación es la misma que nos salía al analizarla por separado anteriormente, sólo que al revés. Recordemos que la habíamos cambiado de signo para facilitar su interpretación.

Respecto a Y_2, vemos que las variables que la hacen crecer son la tasa de aprobados y el número de aprobados. Las demás casi no influyen en ella.

Una vez que hemos terminado con el análisis de las componentes principales, llega el momento de hacer uso de nuestra nueva herramienta gráfica PCA-Gapminder.

Análisis gráfico en PCA-Gapminder

Introducimos las columnas con las puntuaciones de la primera y la segunda componente principal juntas con los datos que contenía *data* en el nuevo objeto *newdat*. Habiendo cargado la librería de googleVis previamente, ya podemos llamar a la función

$$grafico <- gvisMotionChart(newdat, idvar = "ASIGNATURA", timevar = "YEAR")$$

donde *ASIGNATURA* es la variable identidad, los individuos de la muestra; y la variable temporal avanzará de año en año.

Al representar esta función haciendo uso del comando *plot*, se abre una pestaña en el navegador vinculado donde podemos ver la ejecución. Para hacer el estudio representamos en el eje horizontal la primera componente principal y en el eje vertical la segunda componente principal. Con el color representaremos el curso al que pertenecen las asignaturas: azul las de primero y rojo las de segundo. Finalmente, con el tamaño se representará el número de matriculados en la asignatura.

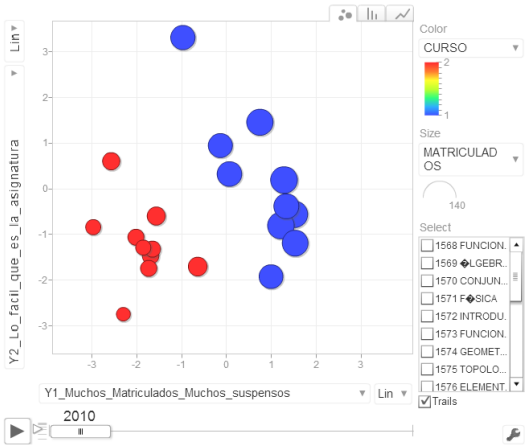


Figura 3.3: Gráfico de las dos primeras componentes de los individuos en 2010.

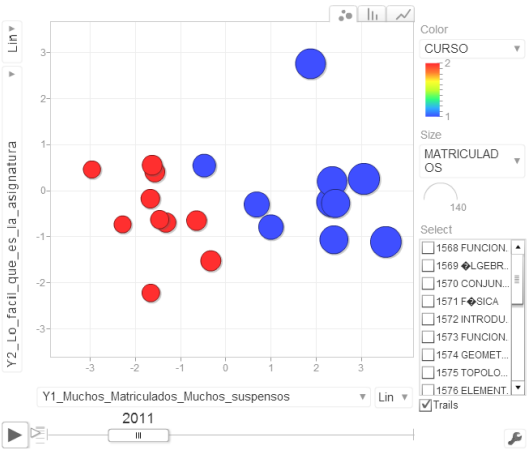


Figura 3.4: Gráfico de las dos primeras componentes de los individuos en 2011.

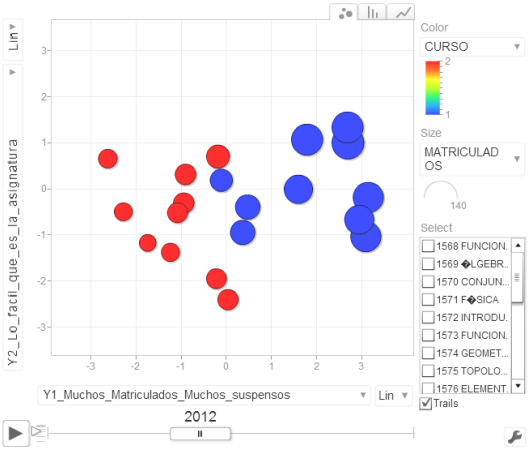


Figura 3.5: Gráfico de las dos primeras componentes de los individuos en 2012.

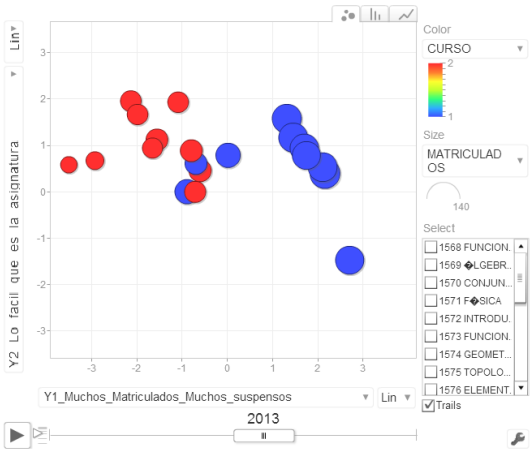


Figura 3.6: Gráfico de las dos primeras componentes de los individuos en 2013.

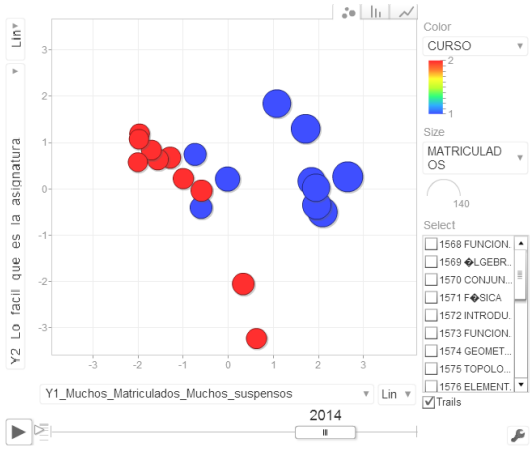


Figura 3.7: Gráfico de las dos primeras componentes de los individuos en 2014.

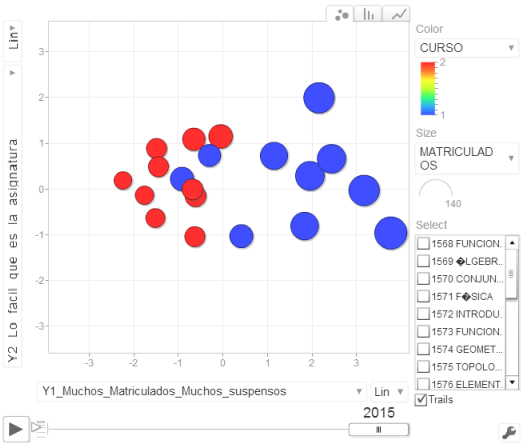


Figura 3.8: Gráfico de las dos primeras componentes de los individuos en 2015.

Interpretación gráfica

El intervalo de tiempo que se analiza comprende desde 2010 hasta 2015. Recordemos el significado aproximado que habíamos dado a Y1 e Y2 para poder interpretar los resultados con más fluidez.

$Y_1 =$ *Asignaturas con muchos matriculados, muchos suspensos y notas bajas*

$Y_2 =$ *Lo "fácil" que es una asignatura*

Por otro lado, el tamaño de las burbujas representa el número de matriculados que hay en la asignatura. El color, el curso al que pertenecen.

Azules: Asignaturas de primer curso

Rojas: Asignaturas de segundo curso

Es obvio que la concentración cromática se mantiene todos los años; asignaturas de primero en el lado derecho y las de segundo a la izquierda. La primera lectura de los datos es, quizá, previsible: en primero muchos alumnos matriculados y un número de suspensos elevado, con notas medias bajas; son los valores positivos en Y1; por el contrario, encontramos valores negativos en Y1 para las asigntauras de segundo curso, entendiendo que el número de matriculas desciende (abandono de algunos alumnos) y el número de suspensos decrece también.

Pero nuestra aplicación pretende descubrir información no previsible, avanzar en los casos singulares y, si es posible, traducir a imagen ciertas relaciones de interdependencia.

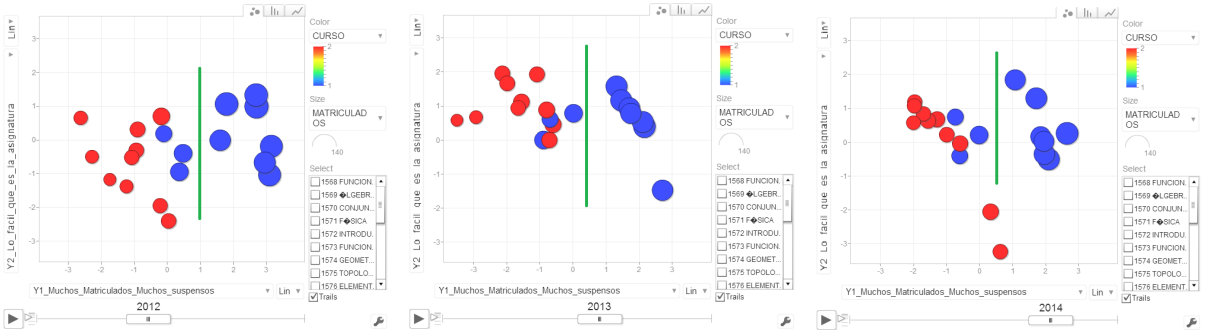
Centremos nuestra atención, por ejemplo, en la dispersión de Y2. En 2010 y 2011 hay una gran dispersión, mientras que en 2013 y 2014 la dispersión disminuye considerablemente. Podríamos afirmar que en el primer periodo los resultados son dispares, mientras que en el segundo tramo temporal las asignaturas tienden a resultados más homogéneos, puede atribuirse a cambios en el perfil del alumnado o a una unificación de criterios por parte de los órganos de coordinación de la facultad.

Otro aspecto interesante es la observación del cuadrante inferior izquierdo que refleja valores inferiores a 0 para las asignaturas de segundo. En el curso 2010 todas las asignaturas de segundo, a excepción de una, se encuentran en valores negativos. Es la radiografía de un curso con calificaciones muy bajas. Sin embargo, 2013 presenta un panorama diferente. Todas, a excepción de una, se sitúan por encima de 0. Es el curso con resultados más positivos.

Sacamos también una reflexión sobre las asignaturas de primero, las burbujas azules. Se agrupan en dos bloques, un bloque de tres asignaturas, que siempre están más a la izquierda (especialmente en los años 2012-2015) y el otro bloque se identifica con el resto de asignaturas, situadas siempre más a la derecha. Podemos observarlo en las imágenes que se muestran abajo, con una línea que separa los dos bloques a los que nos hemos referido.

El bloque que está más a la izquierda significa que son asignaturas con menos matriculados y menos suspensos que las de la derecha. Esas tres asignaturas corresponden a las de formación complementaria, de contenido no estrictamente matemático. Tienen menos matriculados que

las demás de primero porque tienen pocos suspensos y, por tanto, normalmente los alumnos no repiten la asignatura, mientras que las del otro bloque sí.



También resulta interesante identificar una burbuja y seguirle el rastro. Hay algunas que se mueven y se posicionan en lugares muy "raros" respecto a las demás. Veamos alguna.

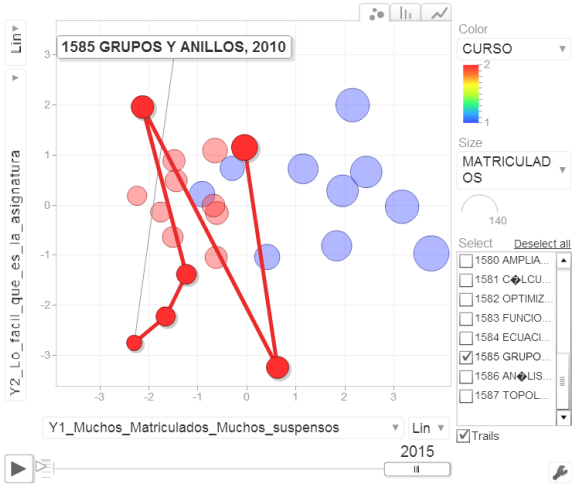


Figura 3.9: Rastro de la asignatura Grupos y Anillos en el periodo 2010-2015.

En la Figura 3.9 se observa el comportamiento de *"Grupos y Anillos"*. Esta asignatura es un ejemplo de las que sufren cambios considerables a lo largo de los años. En los primeros años de grado (2010, 2011, 2012) la asignatura era difícil de aprobar (abajo en Y2) y las notas medias eran mejores (izquierda en Y1), probablemente porque había muchos alumnos no presentados. Sin embargo, en 2013 la situación cambia por completo; posicionándose en la parte positiva del eje vertical, creciendo la tasa de aprobados y considerándose una asignatura "fácil" en ese curso. En 2014 vuelve a producirse otro cambio brusco, se vuelve más difícil de aprobar, aumenta el número de matriculados, el de suspensos y la nota media baja. Finalmente, en 2015, vuelve a cambiar al eje positivo en Y2. Destaca su comportamiento respecto de las otras asignaturas de segundo curso.

Aunque no se aborda el análisis desde esta perspectiva, el color también puede identificar el área o la rama de las matemáticas a la que pertenece cada asignatura. Una interpretación en función de este criterio también puede derivar en atractivas interpretaciones.

3.2. Caso 2. Comunidades Autónomas de España

Para tomar la decisión del objeto de estudio del segundo caso se investigó en varios ámbitos heterogéneos (educativo, económico, demográfico, ...) con el fin de dar un panorama más amplio del uso de PCA-Gapminder. Finalmente optamos por el análisis de diferentes variables en el marco de la distribución territorial española. Por tanto, el objetivo es acercarnos a las realidades de las comunidades autónomas.

Para ello, se solicitó al Instituto Nacional de Estadística (INE) todas las variables que iban a ser necesarias. Este organismo ofrece un servicio electrónico para los ciudadanos. A través de él recopilé todos los ficheros excel de las variables; y una vez hecho esto se creó un fichero único incluyéndolo en conjunto.

Los individuos que vamos a estudiar son las 17 Comunidades Autónomas de España:

Andalucía	Cataluña
Aragón	Comunidad Valenciana
Asturias	Extremadura
Baleares	Galicia
Canarias	Madrid
Cantabria	Región de Murcia
Castilla y León	Comunidad de Navarra
Castilla la Mancha	País Vasco
	La Rioja

Como se comenta anteriormente, las variables se han seleccionado con el fin de poder dar una visión global en distintos ámbitos. Tendremos una variable temporal también que tomará el valor de cada año dentro del periodo 2004-2014 para cada Comunidad Autónoma. A continuación daremos un resumen descriptivo de las variables que se han considerado en el estudio.

- Comunidad Autónoma: Son los "individuos" sobre los que gira el estudio. Mediante el estudio de las componentes principales obtendremos conclusiones acerca de ellas. Son 17.
- Año: Esta variable es anual. Toma los valores entre 2004 y 2014.
- Localización: Esta variable es descriptiva e indica la situación geográfica de cada Comunidad, agrupándolas así en sectores. Puede tomar los valores "Islas","Norte" o "Sur".
- Población: Designa el número de personas que viven en esa Comunidad Autónoma.
- Tasa de nacimiento: Es una medida cuantitativa de la fecundidad en un año. Se calcula dividiendo el número de nacimientos en un año entre la población y multiplicando por 1000. Por tanto, estará expresada en tanto por mil (‰).
- Tasa de defunción: Esta variable se calcula como la proporción entre las personas que mueren respecto al total de la población en un año. Se expresa también en tanto por mil (‰).
- Tasa de extranjeros: Es la proporción de población extranjera residente en la comunidad en relación a la población total. Se expresa en tanto por cien (%).
- Renta media: Esta variable representa la renta anual neta media por persona y unidad consumo. Se mide en euros.

- Tasa de empleo: Indica la tasa de empleo en el tercer trimestre del año de cada Comunidad Autónoma.

- Tasa de médicos: Esta tasa indica el número de médicos a disposición de la comunidad por cada 100000 habitantes.

- Tasa de profesores: Esta tasa indica el número de profesores que hay por cada 100000 habitantes.

- Tasa de condenados: Esta variable indica el número de condenados que hay por cada 1000 habitantes por sentencia firme durante ese año en la comunidad.

Para tener claras cuáles son las variables que intervendrán en el PCA y enumerarlas como X_i a partir de ahora, se muestra el siguiente cuadro síntesis.

X_1	**Tasa de nacimiento**
X_2	**Tasa de defunción**
X_3	**Tasa de extranjeros**
X_4	**Renta anual media por persona**
X_5	**Tasa de empleo**
X_6	**Tasa de médicos**
X_7	**Tasa de profesores**
X_8	**Tasa de condenados**

Cuadro 3.8: Variables numéricas estudiadas.

Comencemos ahora con el estudio de los datos y el PCA mediante RStudio.

Estudio inicial de los datos

Procedemos con los mismos comandos que en el caso de las asignaturas. Introducimos los datos en el fichero mediante *read.table*. Trabajaremos con el objeto *comunidades*, contenedor de todos los datos preparados para analizarse. Si tecleamos *View(comunidades)* podremos visualizar dicho fichero. Como las columnas primera y segunda son los individuos y el año; estas variables se dejan fuera del análisis de componentes principales. Además, dejaremos fuera la variable *Población* y *Localización*, pues la usaremos posteriormente para designar el tamaño y el color de las burbujas respectivamente. Por tanto, en concreto, necesitamos desde la columna 5 hasta la 12 y haremos uso del objeto llamándolo como *comunidades[5:12]*.

Consideramos un resumen inicial mediante *summary(comunidades[5:12])*, donde podemos ver las principales características de las variables por separado.

Tasa.de.nacimiento	Tasa.de.defunción	Tasa.de.extranjeros	Renta.media
Min: 6.216	Min: 6.042	Min: 1.866	Min: 5734
1st. Qu.: 8.738	1st. Qu.: 7.798	1st. Qu.: 5.080	1st. Qu.: 7816
Median: 9.718	Median: 8.755	Median: 9.699	Median: 9171
Mean: 9.751	Mean: 8.809	Mean: 9.865	Mean: 9214
3rd Qu.: 10.869	3rd Qu.: 9.866	3rd Qu. :14.285	3rd Qu.: 10259
Max.: 13.594	Max.: 12.207	Max.: 21.903	Max.: 14312
Tasa.de.empleo	Tasa.de.médicos	Tasa.de.profesores	Tasa.de.condenados
Min: 37.31	Min: 345.0	Min: 1158	Min: 1.353
1st. Qu.: 45.65	1st. Qu.: 426.9	1st. Qu.: 1330	1st. Qu.: 3.350
Median: 49.77	Median: 469.2	Median: 1382	Median: 4.140
Mean: 49.87	Mean: 481.5	Mean: 1402	Mean: 4.092
3rd Qu.:54.01	3rd Qu.: 540.7	3rd Qu.: 1465	3rd Qu.: 4.642
Max.: 64.15	Max.: 641.1	Max.: 1776	Max.: 6.668

Cuadro 3.9: Resumen sobre las variables.

Por otro lado, saber cómo se relacionan entre ellas resultará útil para el posterior PCA, pues dichas relaciones se plasmarán en las componentes principales. Mediante *plot(comunidades[5:12])* obtenemos varios gráficos que muestran las correlaciones entre las variables.

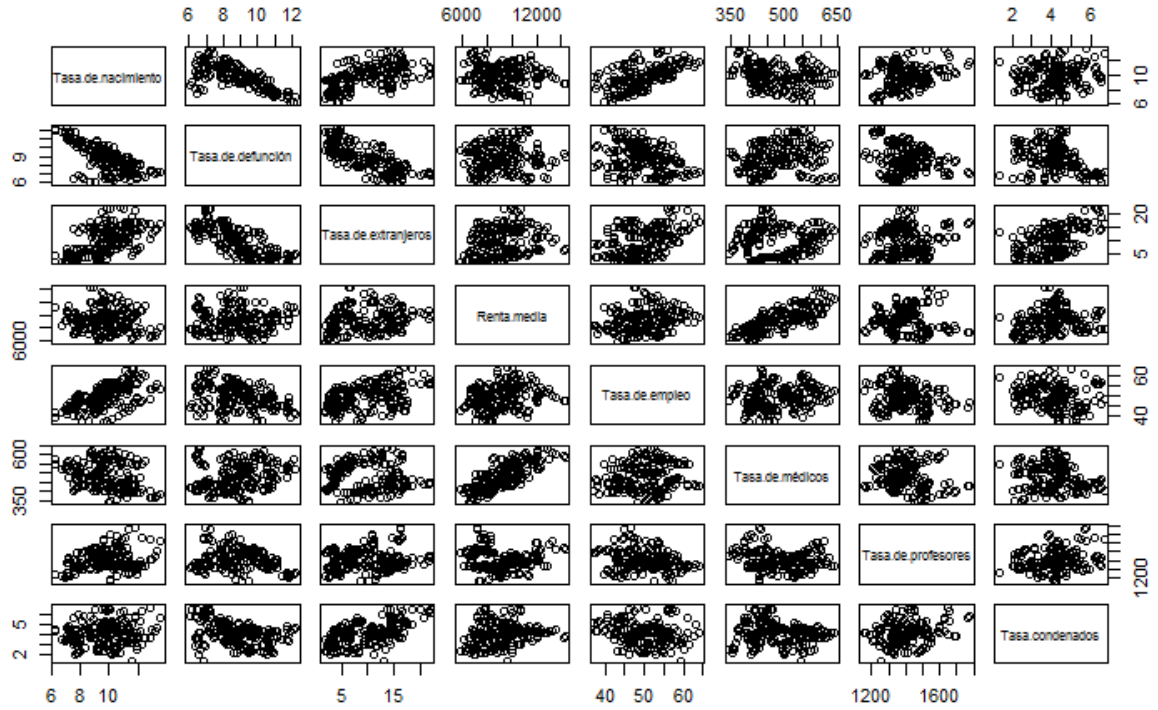


Figura 3.10: Gráficos bidimensionales para las 8 variables.

Se observan algunas claras relaciones entre algunas de ellas. Por ejemplo, se ve la existencia una relación negativa entre *Tasa de nacimiento* y *Tasa de defunción*; mientras que hay relaciones positivas entre *Renta media* y *Tasa de médicos*, o entre *Tasa de extranjeros* y *Tasa de empleo*.

Cálculo de las componentes principales

Para hacer el PCA utilizaremos de nuevo la matriz de correlaciones, con la cual las variables se estandarizan. A pesar de que todas las variables (excepto una), están expresadas en tasas, hay variables con escalas muy diferentes. Véase por ejemplo en el cuadro 3.9 la diferencia tan grande existente entre la mediana de la variable *Renta media*, 9171, frente a la de *Tasa de condenados*, 4.140. Además, algunas tasas están expresadas en tanto por mil y otras en tanto por cien mil, para adecuar los datos a las características de la variable.

A continuación hacemos el PCA, con

$$PCA < -princomp(comunidades[5:12], cor = TRUE)$$

y con *summary (PCA, loadings=TRUE)* examinamos la desviación estándar, la proporción de la varianza y la proporción acumulada (explicado su significado en detalle en el caso 1). Con el PCA se obtienen hasta ocho componentes principales, sin embargo las últimas apenas aportan información. Por esto, sólo se muestra la información correspondiente hasta la quinta componente principal.

Importance	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Standard deviation	1.7132539	1.3966873	1.1995439	0.9401629	0.56338563
Proportion of Variance	0.3669049	0.2438419	0.1798632	0.1104883	0.03967542
Cumulative Proportion	0.3669049	0.6107468	0.7906100	0.9010983	0.94077372

Cuadro 3.10: Importancia de las componentes principales.

Lo más importante que hay que juzgar en este cuadro es la cantidad de información que conservan las componentes. Esto se observaba a través de las proporciones acumuladas (*Cumulative Proportion*). La primera componente principal conserva un 36,66 % de la información inicial, y la primera y la segunda componentes juntas conservan un 61,07 %. Como hay muchas variables en este estudio no bastará con tomar las dos primeras. Tomaremos hasta la cuarta componente principal, conservando así un 90,10 % de la información y haciendo un estudio más completo del problema.

Calculamos las cargas (loadings) con *PCA$loadings* y obtenemos:

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Tasa.de.nacimiento	-0.496610827	0.03731261	0.254622887	0.29470376
Tasa.de.defunción	0.527917215	-0.03571611	-0.005553768	0.03276857
Tasa.de.extranjeros	-0.506365602	-0.14910042	-0.126368288	-0.26751664
Renta.media	-0.001643574	-0.64502167	-0.275605621	0.07631999
Tasa.de.empleo	-0.300819407	-0.35385170	0.510100500	0.04407257
Tasa.de.médicos	0.162341391	-0.63301031	-0.128293845	0.16257833
Tasa.de.profesores	-0.169203547	0.17534636	-0.440165342	0.79136707
Tasa.condenados	-0.269806215	0.04887554	-0.610574795	-0.42432309

Cuadro 3.11: Cargas (loadings) de las componentes principales.

Análisis de las componentes principales

Vamos a analizar el significado de las componentes principales. Como decíamos anteriormente, vamos a seleccionar las primeras cuatro componentes para hacer el estudio. Vayamos ahora estudiando las cargas de cada variable y determinando así una interpretación para cada una. En todos los casos las variables X_i utilizadas para el cálculo de las componentes están estandarizadas[3].

La primera componente principal se calcula como [4]

$$Y_1 = -0{,}496X_1^* + 0{,}527X_2^* - 0{,}506X_3^* - 0{,}001X_4^* - 0{,}300X_5^* + 0{,}162X_6^* - 0{,}169X_7^* - 0{,}269X_8^*.$$

Como casi todos los coeficientes son negativos, es conveniente cambiar la componente de signo. De esta forma,

$$Y_1 = 0{,}496X_1^* - 0{,}527X_2^* + 0{,}506X_3^* + 0{,}001X_4^* + 0{,}300X_5^* - 0{,}162X_6^* + 0{,}169X_7^* + 0{,}269X_8^*$$

donde destacan como coeficientes más significativos los que acompañan a X_1, X_2 y X_3. Indican que la tasa de inmigración es alta, lo cual concuerda con que la tasa de natalidad sea alta (la inmigración de una comunidad hace que se incrementen los nacimientos). Por otro lado, la tasa de defunción es baja. Parece que esta componente es un indicador de la inmigración.

En la segunda componente principal

$$Y_2 = 0{,}037X_1^* - 0{,}035X_2^* - 0{,}149X_3^* - 0{,}645X_4^* - 0{,}353X_5^* - 0{,}633X_6^* + 0{,}175X_7^* + 0{,}048X_8^*$$

destacan los coeficientes de X_4, X_5 y X_6. En esta componente están representadas comunidades con rentas medias bajas, poco personal sanitario y baja tasa de empleo. Todas ellas son un índice de pobreza. Por facilitar después la interpretación, cambiaremos también esta componente de signo.

$$Y_2 = -0{,}037X_1^* + 0{,}035X_2^* + 0{,}149X_3^* + 0{,}645X_4^* + 0{,}353X_5^* + 0{,}633X_6^* - 0{,}175X_7^* - 0{,}048X_8^*.$$

Ahora el significado será la riqueza de la Comunidad. La tercera componente principal se obtiene como

$$Y_3 = 0{,}254X_1^* - 0{,}005X_2^* - 0{,}126X_3^* - 0{,}275X_4^* + 0{,}510X_5^* - 0{,}128X_6^* - 0{,}440X_7^* - 0{,}610X_8^*.$$

Se advierte en este caso que los coeficientes grandes son los que acompañan a X_4, X_5, X_7 y X_8; más específicamente que la tasa de empleo es alta, hay una baja tasa de profesorado, de criminalidad y la renta anual media es baja también. Interpretamos, por tanto, esta componente como comunidades con tasa de empleo alta pero con empleos que necesitan poca formación y que tienen salarios bajos. Por esto la denominaremos "precariedad en el empleo" (aunque hay que reseñar que también se relaciona con tasas altas de empleo).

Finalmente, en la cuarta componente principal

$$Y_4 = 0{,}294X_1^* + 0{,}032X_2^* - 0{,}267X_3^* + 0{,}076X_4^* + 0{,}044X_5^* + 0{,}162X_6^* + 0{,}791X_7^* - 0{,}424X_8^*$$

[3] Estandarizamos la variable i-ésima haciendo $X_i^* = \frac{X_i - mean(X_i)}{sd(X_i)}$

[4] Los decimales de los coeficientes se han truncado a las milésimas.

son los coeficientes que acompañan a X_7 y a X_8 los más relevantes. Hay una alta tasa de profesorado y una baja tasa de criminalidad. Esta componente deducimos que puede indicar el nivel de formación que tiene la población (al recibir una buena formación hay poca tasa de criminalidad).

Una vez que ya hemos analizado cuáles son las variables más representativas en cada componente principal, podemos unir esos factores, dando un significado más concreto a la componente. A continuación se muestra un cuadro con esa "definición" aproximada que caracteriza a cada componente principal.

$Y_1 =$ *La inmigración en la Comunidad.*

$Y_2 =$ *La riqueza de la Comunidad.*

$Y_3 =$ *La precariedad del empleo en la Comunidad.*

$Y_4 =$ *El nivel de formación de la Comunidad.*

Mediante el comando *biplot(PCA,pc.biplot=TRUE)* resulta un gráfico bidimensional de las dos primeras componentes.

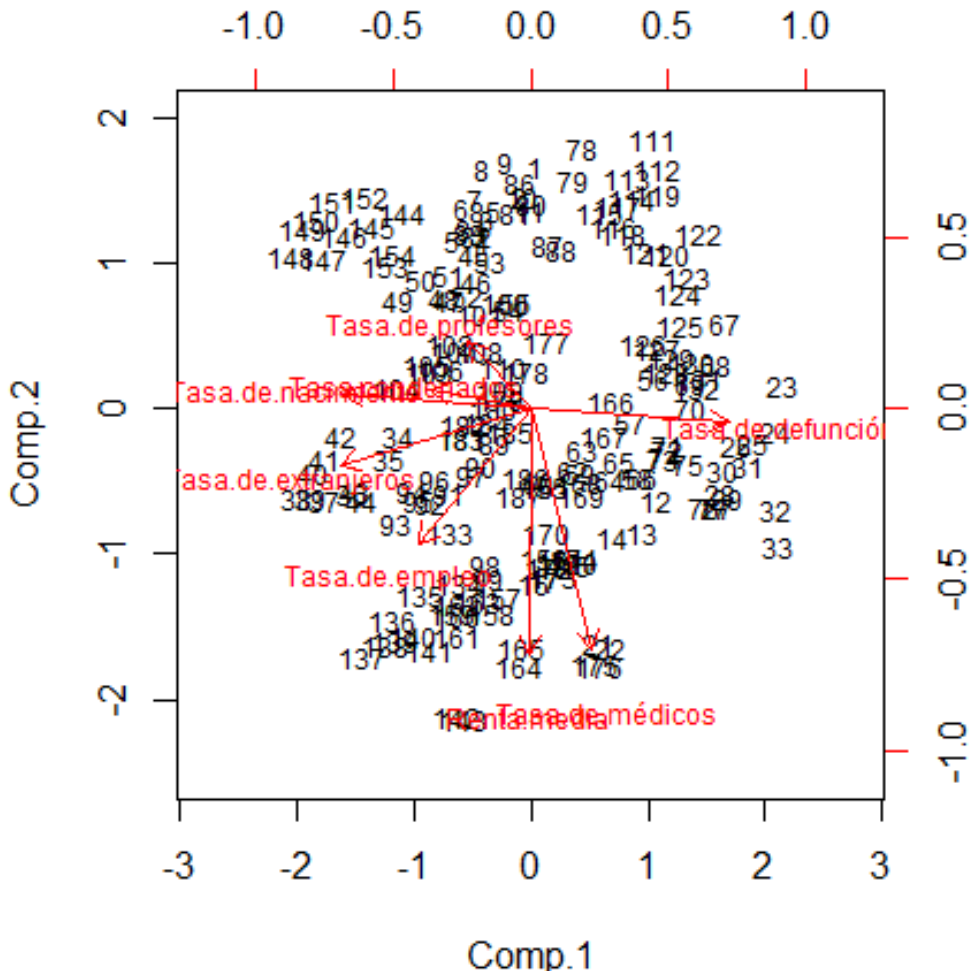


Figura 3.11: Gráfico de las dos primeras componentes principales.

Recordemos que las puntuaciones están representadas en negro etiquetadas con numeración y los vectores rojos son las cargas. Con ayuda de este gráfico podemos también estudiar los significados de las componentes principales; los cuales coincidirán con establecidos en el cuadro anterior. Resulta complicado sacar conclusiones de este gráfico.

Todas las variables están bien representadas pues todas tienen vectores largos, no hay ninguna que se pierda al proyectar. La primera componente se ve incrementada por la variable tasa de defunción; mientras que la tasa de inmigrantes y la de nacimiento la hacen disminuir. El resto de variables no afectan apenas a esta componente. Por otro lado, si nos centramos ahora en la segunda componente vemos que las variables mejor representadas y que más influyen sobre ella son la renta media, la tasa de empleo y la tasa de médicos, todas negativamente. Se podría hacer lo mismo con la tercera y la cuarta componente.

Veamos cómo la nueva herramienta PCA-Gapminder permite realizar un mejor análisis.

Debido a la gran cantidad de gráficos producidos en este estudio (10 años cada periodo y consideración de 4 componentes principales), en lugar de mostrarlos a continuación, irán apareciendo conforme se vayan interpretando en la siguiente sección.

Análisis gráfico en PCA-Gapminder

Al igual que en el caso anterior, almacenamos las puntuaciones de las Comunidades Autónomas para cada componente principal en un objeto, los denotaremos por Y_1, Y_2, Y_3, Y_4. Una vez hecho esto, pegaremos las columnas en el fichero de datos antiguo, creando uno nuevo que llamaremos *new*. En R haremos esto con *cbind(dat,-Y1,-Y2,Y3,Y4)− >new*.

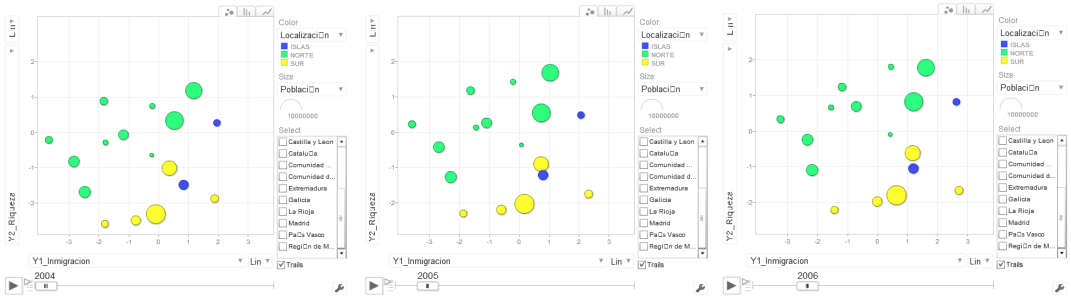
Invocando a la función

grafico < −gvisMotionChart(new, idvar=“Comunidad”, timevar= “Año”),

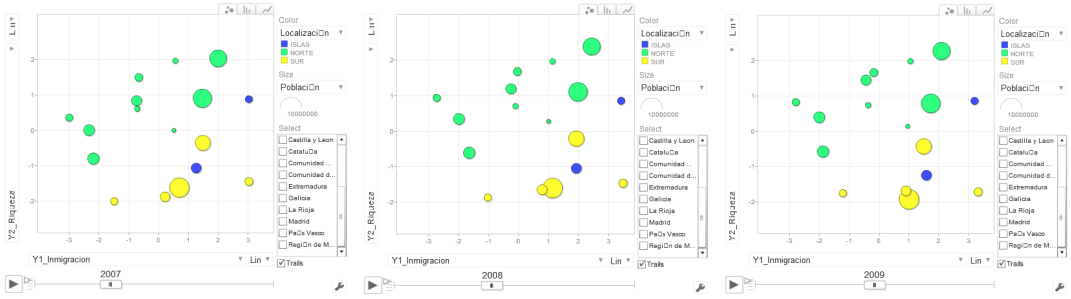
donde *Comunidad* es la variable identidad, la que se representará mediante las burbujas; y la variable temporal avanzará de año en año.

Escribiendo en el script *plot(grafico)*, se ejecuta una pestaña en el navegador donde se obtiene la representación que buscamos. Como en este caso vamos a hacer el estudio combinando las cuatro componentes principales, se irá indicando en cada caso qué componente representamos en cada eje. Con el tamaño representaremos la población que habita en la Comunidad y con el color la localización de la Comunidad Autónoma.

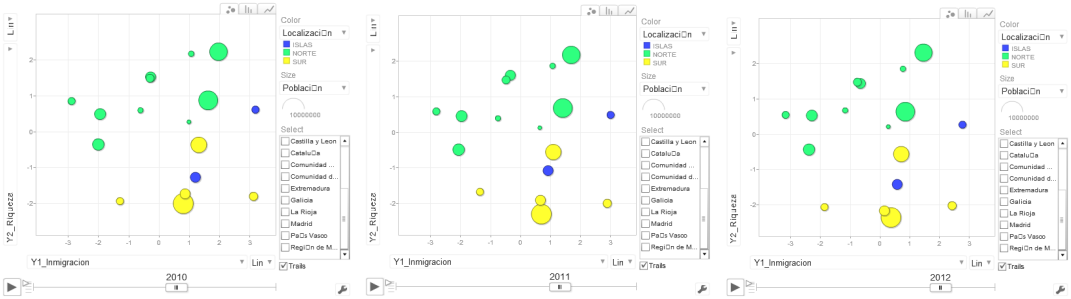
A continuación se muestran capturas de pantalla del periodo temporal completo, año a año. En primer lugar se encuentra el caso de la primera y la segunda componente principal estudiadas, y a continuación las de la tercera y la cuarta componente principal. Posteriormente se dará una interpretación.



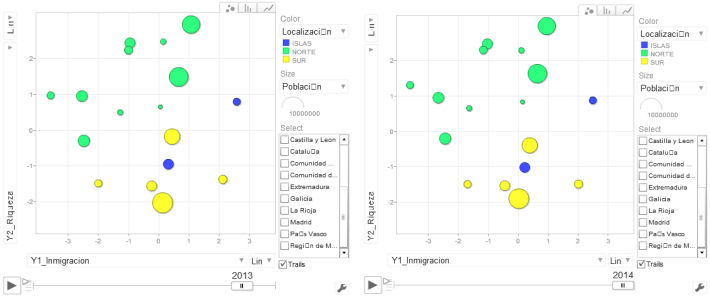
(a) 2004 (b) 2005 (c) 2006



(d) 2007 (e) 2008 (f) 2009

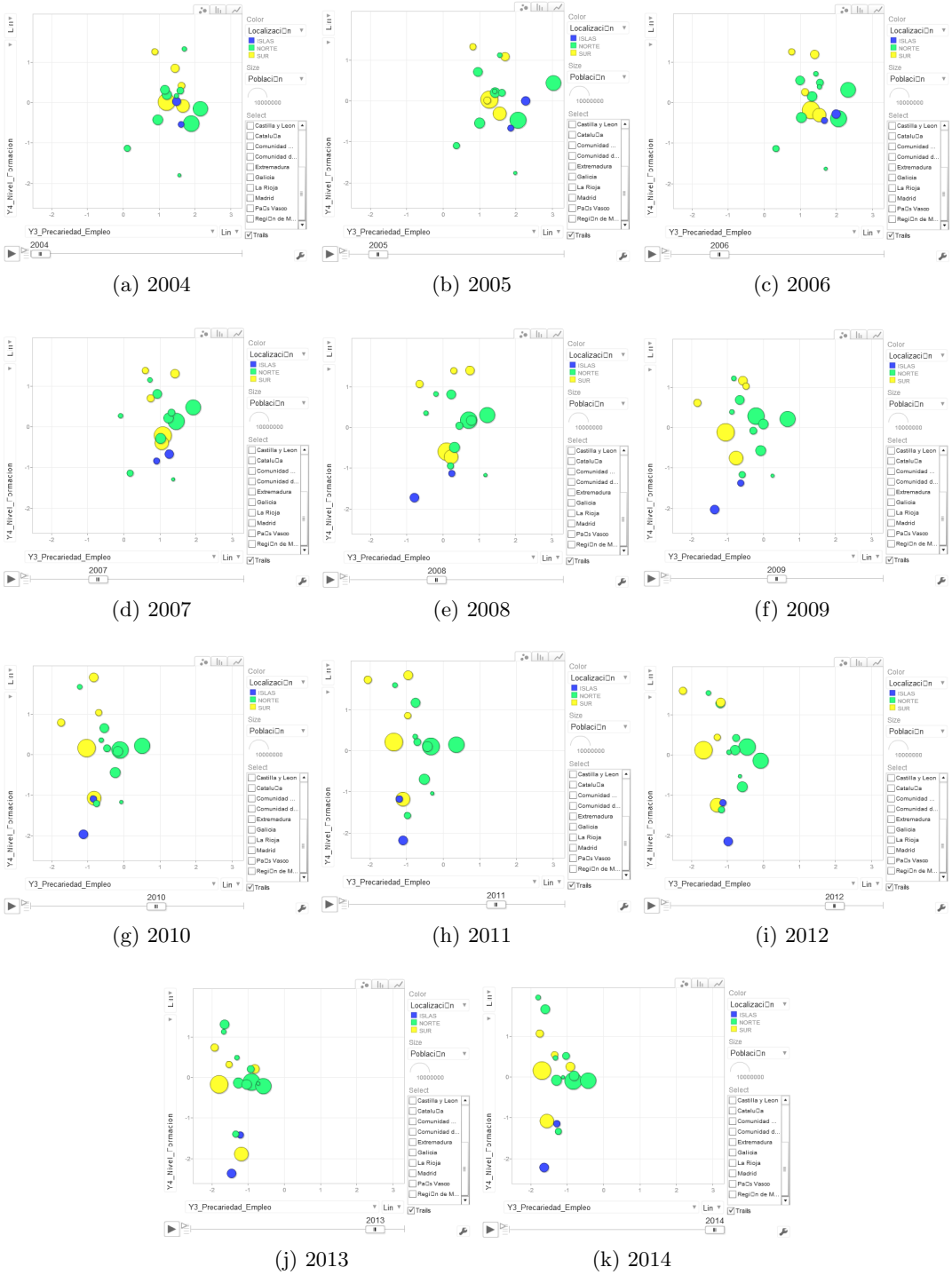


(g) 2010 (h) 2011 (i) 2012



(j) 2013 (k) 2014

Figura 3.12: Capturas del caso de las Comunidades Autónomas. Y1 frente a Y2.



(a) 2004 (b) 2005 (c) 2006

(d) 2007 (e) 2008 (f) 2009

(g) 2010 (h) 2011 (i) 2012

(j) 2013 (k) 2014

Figura 3.13: Capturas del caso de las Comunidades Autónomas. Y3 frente a Y4.

Interpretación gráfica

- Caso 1: Y_1 frente a Y_2.

  Nos proponemos contrastar en este caso los indicadores inmigración y riqueza en las diferentes comunidades autónomas durante el decenio 2004-2014. Norte, sur e islas son los agrupamientos cromáticos realizados. Para la interpretación de estos datos hemos de tener en cuenta esta organización exclusivamente cartográfica, que nos lleva a representar: diez registros en "norte", cinco en "sur" y dos más en "islas".

  Si observamos la dispersión de las burbujas en el gráfico, la primera impresión que obtenemos es que hay una muy diferente realidad en todo el territorio nacional y que esa divergencia se mantiene en el tiempo. En materia de riqueza, una especie de ecuador imaginario parece desgajar el territorio nacional en dos partes: la primera - el norte - se sitúa en valores positivos, en algunos momentos con crecimiento espectacular e insólito (2013); la segunda - el sur - presenta siempre valores negativos y aunque el año señalado se recuperan, no lo hacen lo suficiente como para saltar al otro lado de ese imaginario ecuador de pobreza. La situación insular parece corresponderse en un caso con los valores del norte (Baleares) y en otro con los valores del sur (Canarias). Sin lugar a dudas, se aprecia también el efecto de la crisis de 2008 a 2011 y lo que parece un periodo de recuperación que se marca con claridad en 2013.

  La Región de Murcia presenta, junto a Baleares, la tasa más alta de inmigración. Esta última pertenece al grupo que hemos denominado *insular ecuador norte* y nuestra región al *insular ecuador sur*. Es llamativo, como se ve a continuación en la Figura 3.14, el paralelismo en la evolución de ambas. Observamos que el crecimiento en la inmigración corre paralelo a un incremento en la riqueza. En un determinado momento - que se corresponde con 2010 - la inmigración decrece y también la riqueza. Además, 2012 marca para las dos CCAA idéntica modificación muy positiva de tendencia en riqueza, aunque en este caso sin variación inmigratoria.

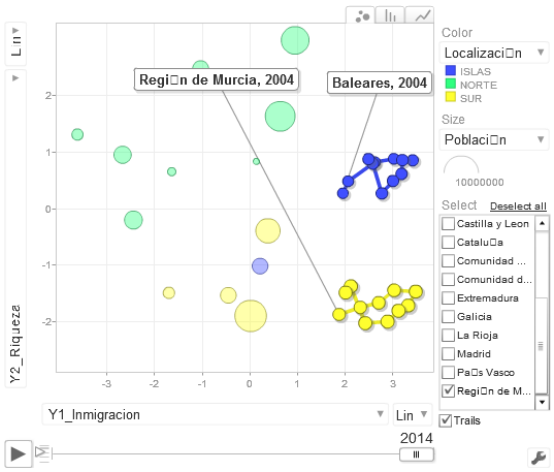


Figura 3.14: Rastro de Baleares y Región de Murcia en las dos primeras componentes en el periodo 2004-2014.

Asturias y Madrid se sitúan en los extremos de la tasa inmigratoria. La primera recoge el valor más pequeño tanto en 2004 como en 2014; la segunda el más alto (si excluimos Murcia y Baleares); ambas presentan índices altos de riqueza, pero si realizamos una sencilla operación podremos comprobar que mientras en el decenio estudiado Asturias gana 1,11 puntos en riqueza, Madrid gana 1,79.

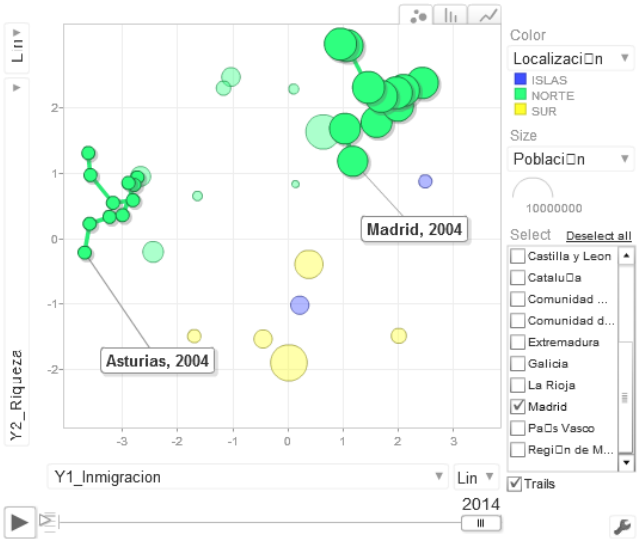


Figura 3.15: Rastro de Asturias y Madrid en las dos primeras componentes en el periodo 2004-2014.

- Caso 2: Y_3 frente a Y_4.

La imagen que contemplamos ahora es completamente distinta. Ni existe dispersión, ni observamos ese ecuador imaginario norte-sur. Por otra parte, las burbujas se desplazan de un extremo a otro respecto al eje horizontal (precariedad en el empleo) casi con "simetría".

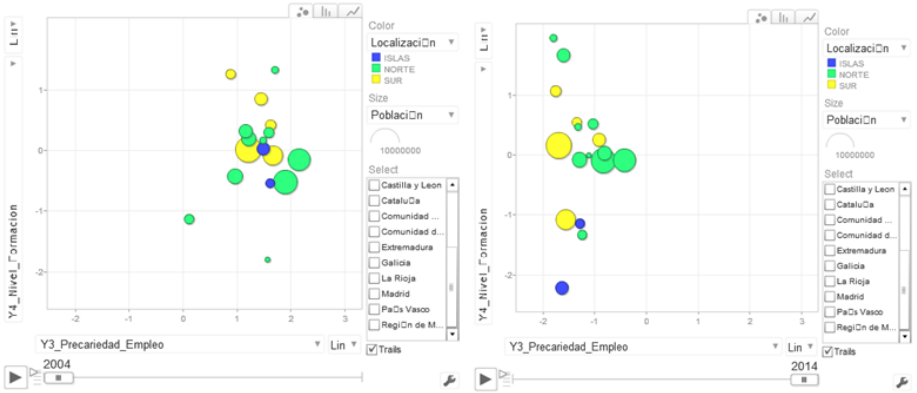


Figura 3.16: Situación en 2004 frente a situación en 2014 respecto a la tercera y cuarta componentes.

Se observa, además, que algunas de las CCAA representadas sufren variaciones muy rápidas o extremas, ascendentes unas, descendentes otras. Una representación, en definitiva, singular, cuya explicación última requeriría estudios complementarios y que aquí nos limitamos a constatar.

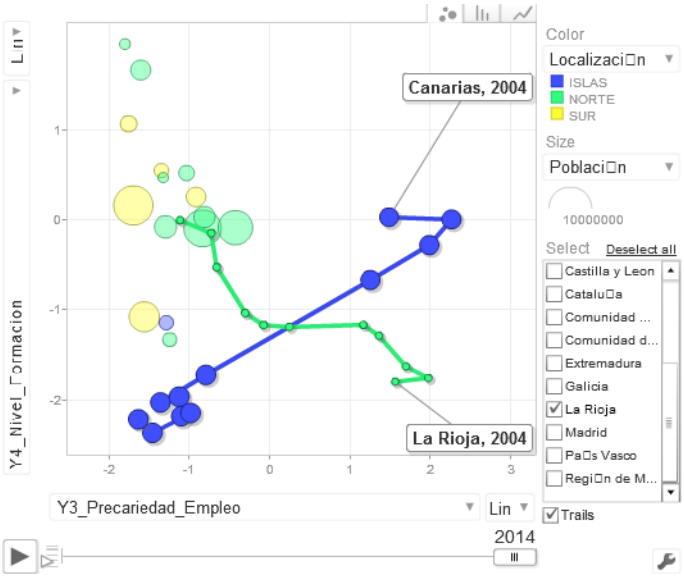


Figura 3.17: Rastro de Canarias y La Rioja en la tercera y cuarta componentes en el periodo 2004-2014.

Este es un caso que nos parece de gran interés por los conceptos representados, aspectos determinantes en el presente y futuro de las sociedades modernas. Son tasas vinculadas a la formación, al empleo y a la inserción normalizada de sus ciudadanos. Que la representación de las diferentes CCAA sea tan desigual es significativo y, aunque sería arriesgado y poco riguroso establecer conclusiones, nos parece un ejemplo muy clarificador de lo valioso que puede ser el instrumento que aportamos en este proyecto fin de grado.

Observemos, en primer lugar, la evolución temporal y, junto al desplazamiento extremo que ya hemos apuntado, llama la atención el ritmo más rápido que se aprecia en todas las comunidades en el eje horizontal durante los años 2007 y 2008. Una de las variables mejor representadas en Y_3 es la tasa de empleo. Coinciden estos años con el inicio de la crisis económica. Podríamos asociar este brusco cambio a valores negativos por la bajada en la tasa de empleo.

Centremos nuestra atención ahora en una de las CCAA que parte de una situación positiva en ambos ejes. Navarra se comporta en el decenio estudiado de la misma forma que el resto en el eje horizontal, pero sigue ocupando los lugares más altos en el vertical; es decir, se suma a la tendencia general en la dimensión "precariedad" y mantiene los más elevados registros de formación.

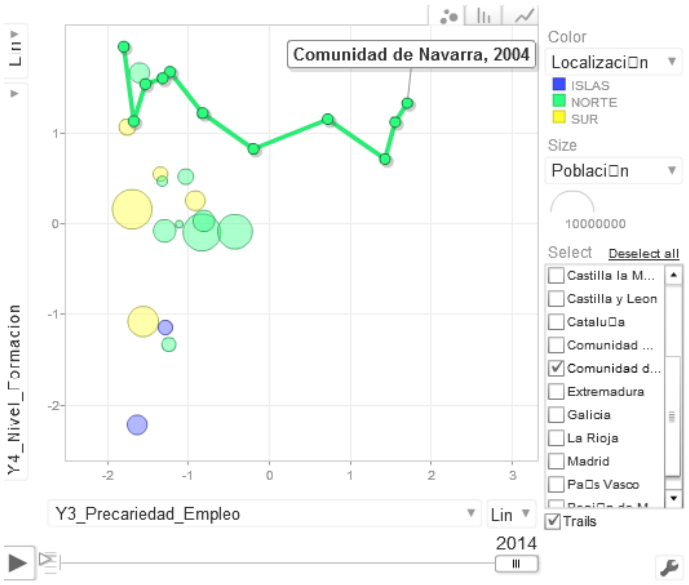


Figura 3.18: Rastro de la Comunidad de Navarra en la tercera y cuarta componentes en el periodo 2004-2014.

Respecto a la evolución de la Región de Murcia, observamos un gran dinamismo que avanza a un ritmo rápido tanto en los valores negativos de la denominada "precariedad laboral" como en los muy positivos designados como "formación".

Región de Murcia	**2004**	**2014**
Precariedad laboral	1.62	-1.75
Nivel de formación	0.43	1.07

Este ritmo veloz es mucho más significativo si lo contrastamos con una CCAA que presenta un comportamiento similar, Castilla La Mancha.

Castilla La Mancha	**2004**	**2014**
Precariedad laboral	1.45	-0.91
Nivel de formación	0.86	0.26

y comprobamos entonces que nuestra Región se mueve a un ritmo todavía más intenso, en torno a un punto más en ambos valores.

Evolución	**Región de Murcia**	**Castilla La Mancha**
Precariedad laboral	-3.37	-2.36
Nivel de formación	+0.64	-0.6

Con estos ejemplos prácticos hemos querido mostrar las excepcionales posibilidades que ofrece PCA-Gapminder, aplicables a muy diferentes ámbitos, así como su utilidad para mostrar realidades no visibles o poco perceptibles en la representación convencional de datos.

Capítulo 4

Conclusiones

1. La estadística es una disciplina esencial, tanto en la vertiente de comprensión del mundo, como en la relativa a las decisiones que lo conforman. Abordar un problema, hoy día, no es una cuestión de escasez de datos; es decir, el hecho de que un problema se solucione o permanezca irresoluble ya no depende de que existan o no datos, sino de que estos sean analizados e interpretados de forma adecuada.

2. Una de las más importantes dificultades para dicha interpretación radica en los procesos en los que intervienen muchas variables; variables que precisan ser analizadas simultáneamente y que se desarrollan mediante técnicas multivariantes. Entre ellas destaca el análisis de componentes principales cuyo objetivo es la reducción de la dimensionalidad mediante la creación de nuevas variables, incorreladas, que son combinaciones lineales de las variables originales.

3. La aplicación Gapminder supuso una revolución al representar de manera original y divulgativa hasta cinco variables, incluyendo dinamismo temporal. En este TFG proponemos combinar los avances de la citada aplicación con el análisis de componentes principales, surgiendo así PCA-Gapminder, que ofrece *datos multivariantes en una nueva visión*, título del presente proyecto.

4. Este proyecto concluye con el logro de presentar una realidad n-dimensional adaptada a la percepción bidimensional y temporal. Por lo tanto, ofrece todas las ventajas de mostrar como un todo lo que en principio constituiría un corpus de datos desagregados.

5. Respecto a posibles líneas de investigación futura, referimos dos: una estrictamente matemática, mediante la profundización en otra técnica multivariante como el análisis discriminante combinado con Gapminder; otra abierta a diferentes disciplinas científicas, ya que como ejemplifican los casos 1 y 2 de nuestro proyecto, son muy útiles, a veces indispensables, este tipo de herramientas matemáticas que, con métodos científicos cualitativos o comparativos, aborden las causas y consecuencias de las complejas realidades que definen el mundo que nos rodea.

Agradecimientos

"Vita brevis, ars longa, occasio praeceps, experimentum periculosum, iudicium difficile."
"La vida es breve, la ciencia eterna, la ocasión fugaz, la experiencia confusa, el juicio difícil."

- A todos los profesores de la Facultad de Matemáticas que me han enseñado durante estos años el significado de esta cita de Hipócrates.

- Gracias, en especial, a D. Jorge Luis Navarro Camacho, y no sólo por su apoyo en este proyecto, sino por transmitirnos, con entusiasmo, la belleza de la estadística.

- También a D. Mathieu Kessler, profesor del Departamento de Matemática Aplicada y Estadística de la Universidad Politécnica de Cartagena, por su ayuda en todo momento.

- A D. Salvador Sánchez Pedreño, por su ayuda en la resolución de dudas relacionadas con el lenguaje Latex.

- Y finalmente, a mi familia.

Bibliografía

[1] EVERITT, B. y HOTHORN, T. *An introduction to applied multivariate analysis with R.* Nueva York: Springer, 2011.

[2] MANCISIDOR, M. *Tack, maestro!* [en línea], «http://www.deia.com». [Última consulta: 20 de mayo de 2017].

[3] NAVARRO, J. L. *Análisis estadístico multivariante usando R.* Apuntes de la asignatura Estadística Multivariante del Grado en Matemáticas de la Universidad de Murcia, 2016.

[4] PEÑA, D. *Análisis de datos multivariantes.* Madrid: McGraw-Hill, 2002.

[5] Contenido Web: «www.ine.es» - *Instituto Nacional de Estadística* - [Última consulta: 13 de marzo de 2017].

[6] Contenido Web: «www.gapminder.org» - *Gapminder* - [Última consulta: 22 de abril de 2017].

[7] Contenido Web: «http://nadandoenunmardedatos.blogspot.com.es» - *Blog de estadística de Mathieu Kessler* - [Última consulta: 28 de abril de 2017].

[8] Contenido Web: «https://www.ted.com» - *TED. Ideas worth spreading.* - [Última consulta: 11 de mayo de 2017].

Anexos

```r
##Ejemplo PCA-GAPMINDER Grado en Matemáticas##
##-------------------------------------------##
library(googleVis)
#library(RJSONIO)
dat<-read.table(file="C:/Users/Inés/Desktop/TFG/Programando/data/datosAsignaturas_copia.csv"
                , dec=",", header=TRUE, sep=";")
View(dat)
summary(dat[5:9])
plot(dat[5:9])
cov(dat[5:9])
M<-cor(dat[5:9])
M
PCA<-princomp(dat[5:9],cor=TRUE)
summary(PCA,loadings=TRUE)
PCA$loadings->T
T[,1]
T[,2]
biplot(PCA,pc.biplot=TRUE, cex=0.8, xlim=c(-4,3))
PCA$scores->S
-S[,1]->Y1_Muchos_Matriculados_Muchos_suspensos
S[,2]->Y2_Lo_facil_que_es_la_asignatura
Y1_Muchos_Matriculados_Muchos_suspensos

biplot(PCA,pc.biplot=TRUE)
cbind(dat,Y1_Muchos_Matriculados_Muchos_suspensos,Y2_Lo_facil_que_es_la_asignatura)->new
View(new)
grafico<-gvisMotionChart(new,idvar="ASIGNATURA",timevar="YEAR", xvar="Y1_Muchos_Matriculados_Muchos_suspensos",
                         yvar="Y2_Lo_facil_que_es_la_asignatura", colorvar="CURSO", sizevar="MATRICULADOS")
plot(grafico)
```

Figura 4.1: Script del código para el caso de las asignaturas del Grado en Matemáticas.

```
1
2  ##Ejemplo PCA-GAPMINDER Comunidades autónomas##
3  ##---------------------------------------##
4  library(googleVis)
5  comunidades<-read.table(file="C:/Users/Inés/Desktop/TFG/Programando/data/datosComunidades_copia.csv",
6                          dec=",", header=TRUE, sep=";")
7  View(comunidades)
8  summary(comunidades[5:12])
9  plot(comunidades[5:12])
10 cov(comunidades[5:12])
11 M<-cor(comunidades[5:12])
12 M
13 PCA<-princomp(comunidades[5:12],cor=TRUE)
14 summary(PCA,loadings=TRUE)
15 PCA$loadings->T
16 T[,1]
17 T[,2]
18 T[,3]
19 T[,4]
20 biplot(PCA,pc.biplot=TRUE, cex=0.75, xlim=c(-2.8,2.8),ylim=c(-2.5,2))
21 PCA$scores->S
22 -S[,1]->Y1_Inmigracion
23 -S[,2]->Y2_Riqueza
24 S[,3]->Y3_Precariedad_Empleo
25 S[,4]->Y4_Nivel_Formacion
26 biplot(PCA,pc.biplot=TRUE)
27 cbind(comunidades,Y1_Inmigracion,Y2_Riqueza,Y3_Precariedad_Empleo,Y4_Nivel_Formacion)->new
28 View(new)
29 grafico<-gvisMotionChart(new,idvar="Comunidad",timevar="Año",xvar="Y1_Inmigracion",yvar="Y2_Riqueza",
30                          sizevar = "Población")
31 plot(grafico)
32
```

Figura 4.2: Script del código para el caso de las Comunidades Autónomas.