Secuencia de aprendizaje para los niveles de entendimiento en la reducción de datos mediante Componentes Principales

José Luis González Bucio*, Graciano Calva Calva, José Manuel Carrión Jiménez, Víctor Hugo Delgado Blas, Joel Omar Yam Gamboa, Walter Magaña Landero, Juan Carlos Avila Reveles.

Universidad de Quintana Roo. Depto. Ingeniería Ambiental, Av. Boulevard Bahía, S/N. Col. el Bosque, C.P. 77019. Tel. 01(983) 8350300, Chetumal, Quintana Roo. Centro de Investigación y de Estudios Avanzados unidad Zacatenco. Departamento de Biotecnología, Av. Instituto Politécnico Nacional No. 2508. Col. San Pedro Zacatenco, C.P. 07360, Ciudad de México, México. *contacto: buciojos@uqroo.edu.mx

Resumen

El volumen de datos generados en todos los aspectos, ya sean laborales y no laborales, complican su interpretación y ha traído como consecuencia la necesidad de utilización de métodos especiales del procesamiento de los datos. Por ejemplo, en el análisis ambiental que está conformada por una serie de métodos matemáticos, estadísticos y químicos, permiten el reconocimiento de relaciones complejas y patrones de comportamiento ocultos en grandes cantidades de datos, en una amplia variedad de situaciones en diversos campos (Einax, 1992). Los resultados de las mediciones medioambientales se caracterizan usualmente por su elevada variabilidad, en la que se incluye, además de su variabilidad natural, la incertidumbre resultante de los procesos analíticos como el muestreo, la conservación de las muestras, las mediciones analíticas, etc., de manera que al emplear métodos univariados para su estudio se pierde a menudo información (Grupo de Quimiometría, 2001). Una de las principales herramientas para esta complejidad de datos es el Análisis por Componentes Principales (ACP), esta ha sido empleada en los últimos 20 años, en la interpretación de datos numéricos resultantes de los diseños de experimentos y de la medición de numerosas variables simultáneamente. El análisis de componentes principales (ACP) es una técnica multivariante diseñada en 1901 por Karl Pearson con la finalidad de encontrar líneas y planos que mejor se ajusten a una nube de puntos en el espacio. La técnica de Pearson ha sido modificada a través de los años (Cela, 1994) empleándose actualmente en muchos tipos de problemas multidimensionales (Quintana., 1994; Carlosena, 1999; Cal, 2001).

Introducción

El análisis de componentes principales (ACP) es una técnica multivariante diseñada en 1901 por Karl Pearson con la finalidad de encontrar líneas y planos que mejor se ajusten a una nube de puntos en el espacio. La técnica de Pearson ha sido modificada a través de los años (Cela, 1994) empleándose actualmente en muchos tipos de problemas multidimensionales (Quintana., 1994; Carlosena, 1999; Cal, 2001). La idea básica del ACP es encontrar un pequeño número de combinaciones lineales no correlacionadas a partir de las n variables originales, las cuales expliquen el mayor por ciento de la variabilidad total de los datos originales. Es decir, pasar el conjunto de variables primitivas a un nuevo espacio, trasladando la máxima información original contenida en el espacio multidimensional analizado, a un espacio de dimensionalidad reducida, como por ejemplo tri o bidimensional (Massart D.L.,

1998) Los estudios multivariados, en cualquier campo al que se apliquen, comienzan frecuentemente a partir de la construcción de una matriz de correlación. Esta constituye una tabla simétrica de coeficientes de correlación de cada variable respecto a cada una de las otras; de ellas pueden emerger patrones y estructuras que a menudo no son distinguibles por simple inspección de los datos. El análisis de correlación sirve como técnica primaria descriptiva estimadora del grado de asociación entre las variables involucradas en el estudio (Sokal, 2012). Las variables analizadas deben presentar cierto grado de correlación, o sea, que aporten relativamente el mismo tipo de información (por ejemplo, un conjunto de distintos parámetros de composición metálica de sedimentos y organismos) para que pueda llevarse a cabo la reducción del número de variables, manteniendo un porcentaje elevado de la información que aportaban las variables originales (Quintana, 1994). Por ello, primeramente debe comprobarse la adecuación de los datos para la realización del ACP mediante la prueba de esfericidad de Bartlett (Cela, 1994) que revela cuándo la matriz de correlación es una matriz de identidad, lo cual podría indicar que las variables no estuvieran correlacionadas.

En este trabajo distinguiremos entre el Análisis Exploratorio de Datos (AED) para datos univariados (AEDU), y el AED para datos multivariados (AEDM).

Metodología

Primeramente, se realizó el diseño, implementación y evaluación de una secuencia de aprendizaje, para el Análisis de Componentes Principales (ACP). Se impulsó en el estudiante una compresión intuitiva de la reducción de datos para el (ACP) y se desarrolló en el estudiante la compresión del concepto de Componente Principal y finalmente, se aplicó e interpretó el (ACP).

Las aplicaciones del (ACP) son numerosas y entre ellas podemos citar la clasificación de individuos, la comparación de poblaciones, la estratificación multivariada, etc. El Análisis de Componentes Principales (ACP) es una técnica cuyo objetivo principal es hallar combinaciones lineales de variables representativas de ciertos fenómenos multidimensionales, con la propiedad de que exhiban varianza mínima y que a la vez no estén correlacionadas entre sí. Para obtener tales combinaciones es necesario construir la matriz de varianzas y covarianzas de esas variables (Dallas. E.J., 2000).

Permite reducir la dimensionalidad de los datos, transformando el conjunto de p variables originales en otro conjunto de q variables no correlacionadas $(q \leq p)$ llamadas componentes principales. Las p variables son medidas sobre cada uno de los n individuos, obteniéndose una tabla de datos o matriz de datos de orden np $(p < n)$. La varianza de la primera componente mientras mayor sea su varianza, mayor será la cantidad de información en dicha componente. Por ello las sucesivas combinaciones o variantes de las componentes se ordenan en forma descendente de acuerdo a la proporción de la varianza total presente en el problema, que cada una de ellas explica (Pérez C., 2000).

Generación de las Componentes Principales desde los datos.

La reducción de datos mediante componentes principales esencialmente consiste en organizar los datos en una matriz X de tamaño $(n \times p)$ donde (p) es número de variables en el estudio y n es el número de observaciones. Las variables se denotan como $X_1, X_2, \dots, X_p$, de modo que:

$$X = \begin{pmatrix} x_{11}\, x_{12} \dots . x_{1p} \\ x_{21}\, x_{22} \dots . x_{2p} \\ \dots \dots \dots \dots \dots \\ x_{n1}\, x_{n2} \dots x_{np} \end{pmatrix}$$

La matriz de varianzas-covarianzas muestral sería la matriz $S = (Sjk)$ de tamaño $p \times p$ donde:

$$S_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k) \quad j, k = 1,2, \dots, p$$

La primera componente principal estaría en la dirección que capture la máxima variabilidad de los datos originales. Esto se consigue resolviendo la ecuación:

$$(S - l_{(1)} I)\underline{a}_{(1)} = \underline{0}$$

Donde $l_{(1)}$ es el multiplicador de Lagrange. El valor de $l(1)$ se escoge de modo que:

$$|S - l_{(1)} I| = 0$$

Aquí $l_{(1)}$ es el máximo eigenvalor de la matriz de varianzas-covarianzas S, y $a(1)$ es su correspondiente eigenvector. El procedimiento para la obtención de las demás direcciones es directo (Dillon y Goldstain, 1984).

Para efectos de adaptar la propuesta de Garfield y Ben-Zvi (2005) y las de Van Hiele en la construcción de nuestra secuencia de aprendizaje, primeramente, describimos los postulados centrales del modelo de Van Hiele:

1. Se pueden encontrar varios niveles de perfección en el razonamiento de los estudiantes de Matemáticas (Estadística).

2. Un estudiante solo podrá comprender realmente aquellas partes de las Matemáticas (Estadística) que el profesor le presente de manera adecuada a su nivel de razonamiento.

3. Si una relación matemática (concepto estadístico) no puede ser expresada en el nivel actual de razonamiento de los estudiantes, será necesario esperar a que éstos alcancen un nivel superior de razonamiento para presentársela.

4. No se puede enseñar a razonar a una persona de una determinada forma. Pero sí se le puede ayudar, mediante una enseñanza adecuada de las Matemáticas (Estadística), a que llegue a razonar de esa forma.

Con base en los postulados anteriores, se determinan las características de cada uno de los niveles de comprensión de la reducción de datos:

Nivel 1. Desarrollo intuitivo de la Reducción de Datos

• El estudiante es capaz de reconocer que "muchas" variables no son fáciles de manejar y de que no es fácil darles sentido o interpretarlas globalmente.

• El estudiante puede explicar, por qué un conjunto de variables están correlacionadas y puede determinar diversos niveles de correlación.

Nivel 2. Implementación de la Reducción de Datos

• El estudiante puede utilizar una matriz de gráficas bivariadas para identificar posibles asociaciones entre dos variables.

• El estudiante cuenta con habilidad para interpretar la información de una matriz de varianzas y covarianzas.

Nivel 3. Interpretación de los resultados de la Reducción de Datos

• El estudiante puede utilizar tanto las gráficas como la matriz de varianzas y covarianzas para detectar posibles redundancias en las variables.

• Al examinar cuidadosamente estos grupos de variables, los estudiantes pueden extraer información sobre el comportamiento del fenómeno del cual provienen los datos.

Nivel 4. Comprensión global de la Reducción de Datos y el ajuste de modelos

• El estudiante es capaz de construir diagramas de dispersión e interpretar la distancia entre los puntos en un sistema de ejes definido por dos variables.

• El estudiante posee habilidad para transformar y reorganizar los datos originales y definir nuevos constructos (grupos de variables), los cuales interpreta en términos del fenómeno de referencia.

Nivel 5. Valoración de la Reducción de Datos como parte del pensamiento estadístico

• El estudiante es capaz de explicar una variable a través de un conjunto de otras variables relacionadas con la primera.

• El estudiante puede explicar el agrupamiento de variables a través de constructos que subyacen a las mismas.

Actividades de la secuencia de aprendizaje

Se integraron 5 equipos en binas, seguidamente se desarrollaron 3 actividades con alumnos del cuarto semestre de la carrera de Ingeniería Ambiental de la División de Ciencias e Ingenierías de la Universidad de Quintana Roo, cada grupo de actividades se desarrollaron durante 2 horas en el aula con: Apertura de trabajo, desarrollo y evaluación de la secuencia de aprendizaje.

Resultados

Informe del desarrollo de la secuencia

Se aplicó la secuencia de aprendizaje a los alumnos del cuarto semestre de la carrera de Ingeniería Ambiental de la División de Ciencias e Ingeniería de la Universidad de Quintana Roo en el ciclo de primavera en el mes de febrero del 2017. Organizándose en 5 equipos de parejas (binas). Al inicio el profesor dio instrucción de la secuencia de aprendizaje por cada uno de los niveles de compresión, los cuales son los siguientes:

Nivel de comprensión 1

Desarrollo intuitivo de la reducción de datos

En esta fase, como primera actividad se les presentó a los alumnos la introducción al caso de estudio (el estudio de un grupo de pacientes hipertensos, con relación a las variables de riesgo de enfermedad coronaria en la ciudad de Chetumal, Quintana Roo. Así como, identificar cuáles son las variables que incrementan el riesgo de presentar una enfermedad coronaria y un posible infarto).

Los alumnos respondieron correctamente, diciendo que se trataba de 7 variables y que se estudiaba un caso poblacional de problemas de hipertensión en pacientes adultos. Solamente un equipo (bina) respondió el valor que puede tomar cada dato. Expresaron que cada dato tenía diferente magnitud y unidad de mediada (kg, mm de Hg, %, pulsaciones por minuto, etc.). Todos los equipos respondieron correctamente, comentaron que eran 20 pacientes o individuos estudiados. Solamente 3 equipos respondieron correctamente cual era el objetivo de medición de los datos en la muestra. Conocer los riesgos de enfermedades coronarias o cardiovasculares. En la pregunta, donde se pide el porcentaje de los pacientes con peso más de 100 kg. Respondieron correctamente 2 equipos de 5 (3 binas contestaron incorrectamente). Al preguntar, ¿intuitivamente crees que es más difícil manejar y analizar los datos con menos o con más variables? 2 equipos respondieron correctamente, los otros 3 equipos comentaron que es más difícil manejar menos variables, aún no tenían los conocimientos necesarios para poder contestar el objetivo principal de esta secuencia de aprendizaje.

En la sección 2 de preguntas relacionadas con la Figura 1 (Histogramas): los 5 equipos (binas) contestaron correctamente, sin embargo, para algunos alumnos fue difícil interpretar los histogramas, por lo que los alumnos tuvieron que apoyarse de las Tablas 1 y 2. En la tercera sección de la actividad 1, la mayoría de los equipos contestaron correctamente y supieron

ubicar la media, el valor mínimo y máximo de cada variable de la Tabla 2 en los Histogramas de la Figura 1.

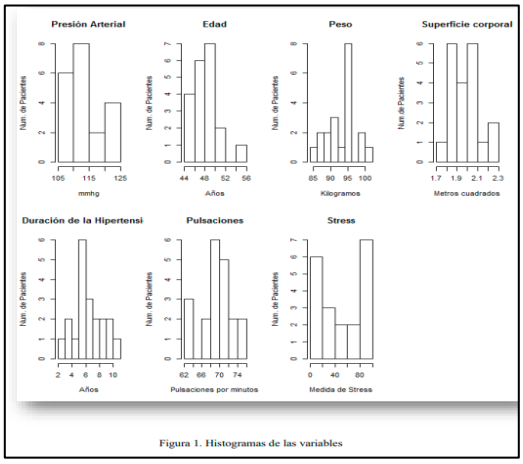

Figura 1. Histogramas de las variables

Tabla1. Tabla de datos

Número de Paciente	Presión_ arterial	Edad	Peso	Superficie corporal	D. Hipertensión	Pulso	Stress
1	105	47	85	1.75	5.1	63	33
2	115	49	94	2.1	3.8	70	14
3	116	49	95	1.98	8.2	72	10
4	117	50	95	2.01	5.8	73	99
5	112	51	89	1.89	7	72	95
6	121	48	100	2.25	9.3	71	10
7	121	49	100	2.25	2.5	69	42
8	110	47	91	1.9	6.2	69	8
9	110	49	89	1.83	7.1	69	62
10	114	48	93	2.07	5.6	64	35
11	114	47	94	2.07	5.3	74	90
12	115	49	94	1.98	5.6	71	21
13	114	50	92	2.05	10	68	47
14	106	45	87	1.92	5.6	67	80
15	125	52	101	2.19	10	76	98
16	114	46	95	1.98	7.4	69	95
17	106	46	87	1.87	3.6	62	18
18	113	46	95	1.9	4.3	70	12
19	110	48	91	1.88	9	71	99
20	122	56	96	2.09	7	75	99

Nivel de comprensión 2

Implementación de la reducción de datos

En esta fase la mayoría de los alumnos ya tenían los conocimientos de: el término de dispersión y correlación (asociación), por lo que pudieron proyectar en las gráficas de dispersión correctamente las variables (presión y peso) de cada paciente. La mayoría de los estudiantes, en este nivel de comprensión ya tenían conocimientos de los conceptos de varianza y covarianza (dispersión), y del concepto de dispersión (asociación) de las variables.

Los alumnos comenzaron a relacionar las variables, algunos alumnos ya descartaban una variable porque veían que el comportamiento de algunas variables eran similares, la mayoría de los alumnos ya tenían los conocimientos de correlación, dispersión, varianza, media, etc. Por ejemplo: algunos alumnos comentaban que si la persona presentaba sobrepeso tendría la presión arterial elevada.

Nivel de comprensión 3

Comprensión global de la reducción de datos y el ajuste de modelos

En esta fase, como tercera actividad se estimó la comprensión global de la reducción de los datos. La mayoría de los alumnos ya tenían los conocimientos de: el término de dispersión y correlación (asociación), por lo que pudieron proyectar en las gráficas de dispersión correctamente las variables (presión y peso) de cada paciente. La mayoría de los estudiantes, en este nivel de comprensión ya tenían conocimientos de los conceptos de varianza y covarianza (dispersión), y del concepto de dispersión (asociación) de las variables.

En el desarrollo de esta actividad, generalmente todos los alumnos ya conocían que era un componente principal, supieron proyectar los datos (puntos) en la recta (componente principal).

Al observar las Figuras que ilustraban los componentes principales identificaron la línea con mayor y menor error, también las Figuras que presentaban mayor correlación, además, relacionaron la correlación con el error en el Componente Principal.

Resumen general

Con la información generada de este primer análisis de resultados, se tuvo un criterio para clasificar cada equipo de estudiantes en dos grupos: destacados y poco destacados. En este sentido se identificaron los siguientes grupos:

Grupo I de alumnos destacados: 3 equipos y, Grupo II de alumnos poco destacados: 2 equipos. Figuras 3, 4,9 y 11.

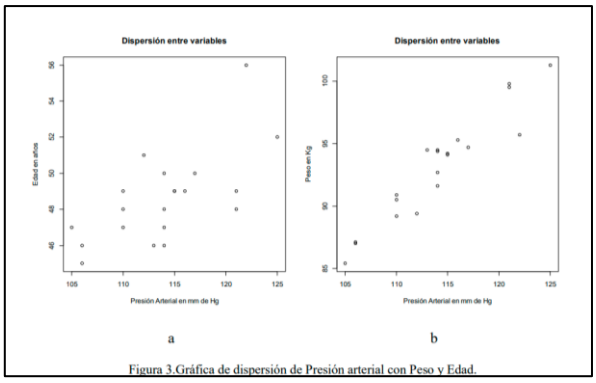

Figura 3.Gráfica de dispersión de Presión arterial con Peso y Edad.

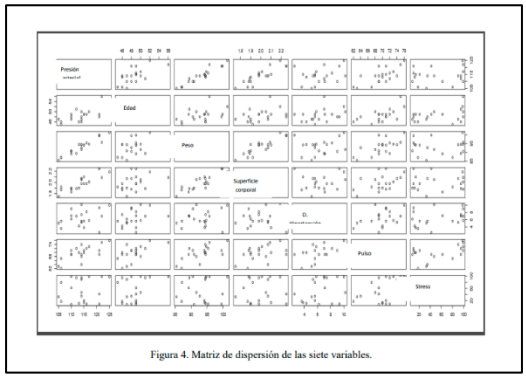

Figura 4. Matriz de dispersión de las siete variables.

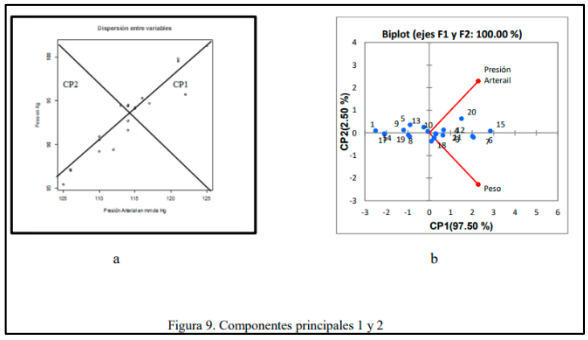

Figura 9. Componentes principales 1 y 2

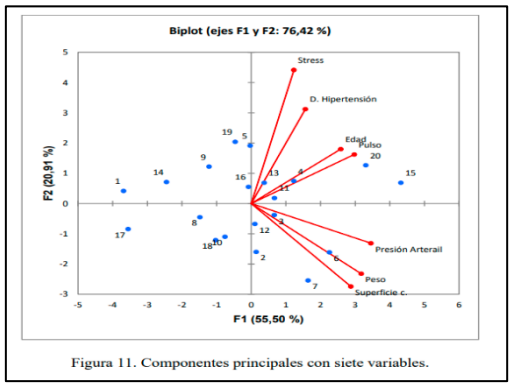

Figura 11. Componentes principales con siete variables.

Conclusiones

El modelo de Van Hiele, abarca el aspecto descriptivo, el cual identificamos en el momento que los alumnos experimentaron diferentes formas de razonamiento, partiendo de lo geométrico y fueron valorando el progreso del entendimiento. ***En el aspecto instructivo***, el profesor a partir de la instrucción de la secuencia de aprendizaje dio las pautas para favorecer el razonamiento geométrico de los datos procesados. En los tres niveles de comprensión

aplicados: 1.- Desarrollo Intuitivo de la de la Reducción de Datos: en este nivel los alumnos difícilmente llegaron al entendimiento de cómo y por qué reducir datos en un análisis estadístico. En el segundo nivel de comprensión: Implementación de la Reducción de Datos, los alumnos alcanzaron a un mejor entendimiento del manejo y reducción de los datos y finalmente, en el tercer nivel de comprensión: Comprensión global de la Reducción de Datos y el ajuste de modelos, los alumnos alcanzaron un mejor entendimiento y desarrollaron sus habilidades, aplicando lo aprendido en las asignaturas algebra lineal y geometría en esta secuencia de aprendizaje, en ese momento pudimos percibir que desarrollaron los alumnos la compresión, aplicación e interpretación del concepto de Componentes Principales.

Citas bibliográficas

Carlosena A. (1999). Classificación of edible vegetables affected by different traffic intensities usin potencial curves. Talanta, 48, 745.

Cela R. (1994). Avances en Quimiometría Práctica. Ed. Univ. Stgo. de Compostela.

Dillon, W.R. y Goldstain, M. (1984). Multivariate A nalysis, Methods and Applications. Wiley, New York.

Einax J.E. (1992). Multivariate data analysis in environmental analytical Chemestry. GIT-FachzLab, 36(8):815.

Garfiel, J. y Ben-Zvi, D (2005). A Framework for Teaching and Assessing Reasoning about Variability. Statistics Education Research Journal, 4(1). 92-99.

Grupo de Quimiometria, i Qualimetria de Tarragona. Quimiometría una disciplina per al ‘analisi quimica. Universistat de Rovira i Virgilia. 2001.

Massart D.L. (1998). Handbook of Chemometrics and Qualimetrics: Part B. Ed. Elsevier Sc.

Pérez C. (2000). Técnicas de análisis multivariate de datos. Aplicaciones con SPSS. Madrid; Pearson, pp. 121-154.

Quintana I., Mora G. (1994) Análisis de suero Humano por Espectroscopía Atómica. Tesis de Diploma, Facultad de Química, Universidad de la Habana, Cuba.

Sokal, R. (2012). Biometría: los principios y la práctica de la estadística en la investigación biológica. 2ª ed. Libro en PDF en la Revista de la Sociedad Estadística Real Seria.