APLICACIÓN DE TÉCNICAS ESTADÍSTICAS MULTIVARIADAS CON EL LENGUAJE DE PROGRAMACIÓN R EN INVESTIGACIONES EDUCATIVAS DEL NIVEL SUPERIOR

Application of multivariate statistical techniques with the R programming language in higher level education research

Débora Chan, Universidad Tecnológica Nacional (FRBA, INSPT), Argentina
debiechan@gmail.com

María Gabriela Galli, Universidad Tecnológica Nacional (INSPT), Argentina
gabriela.galli@inspt.utn.edu.ar

Chan, D. y Galli, M. G. (2020). Aplicación de técnicas estadísticas multivariadas con el lenguaje de programación R en investigaciones educativas del nivel superior. *RAES*, 12(20), pp. 123-136.

Resumen

La estadística es una disciplina universalmente aplicada para describir clasificar, inferir, decidir y cuantificar la probabilidad de error en la toma de decisiones científicas. Su aplicación en estudios cuantitativos y cualitativos ya está consolidada, sin embargo, su utilidad en áreas sociales y específicamente en investigación educativa se ha visto impulsada recientemente por la difusión del software *open source* y el crecimiento exponencial de la ciencia de datos. Con el objetivo de estudiar la contribución diferencial de las técnicas estadísticas multivariadas y del lenguaje R, se han seleccionado los trabajos de investigación educativa del nivel superior de un período reciente. Se analizaron las técnicas multivariadas empleadas en las mencionadas investigaciones y se las comparó con trabajos que analizaban contextos muy similares, pero que no aplicaban estas técnicas. A partir de ello se desprende la evolución de la aplicación de estas técnicas y del lenguaje R en las investigaciones de corte educativo.

Palabras Clave: análisis multivariado/ lenguaje R/ investigación educativa/ educación/ nivel superior.

Abstract

Statistics is a discipline universally applied to describing, classifying, inferring, deciding and quantifying the probability of error in scientific decision-making. Its application in quantitative and qualitative studies is already consolidated; however, its usefulness in social areas, specifically in educational research has recently been boosted by the spread of open source software and the exponential growth of data science. In order to study the differential contribution of multivariate statistical techniques and R language, the educational research recently conducted at the higher has been selected.The multivariate techniques used in in such research were analyzed and compared with works that analyzed very similar contexts, but that did not apply

these techniques. From there on, the evolution of the application of these techniques and of the R language in educational research can be accounted for.

Key words: multivariate analysis/ language R/ educational research/ education/ higher education

Introducción

La investigación educativa tiene como fin fundamental, la comprensión e interpretación de los fenómenos que ocurren dentro del ámbito educativo. Asimismo, pretende brindar aportes con el propósito de mejorar las acciones, intervenciones y resolver problemas de ese campo de estudio. Este tipo de investigaciones se caracteriza por proveer una gran variedad de enfoques y metodologías de acuerdo con la complejidad del objeto de estudio. Según la naturaleza de los datos, las investigaciones trabajan con enfoques cuantitativos, cualitativos o mixtos. De acuerdo con la finalidad que se persigue, Bisquerra Alzina (2009) identifica en este campo los estudios con métodos orientados a obtener conocimiento básico, mediante la descripción, explicación, predicción y comprensión en profundidad de los hechos, todos vinculados a cuestiones prácticas. Asimismo, los que persiguen obtener conocimiento aplicado, cuyo propósito es comprender la realidad sobre datos críticos que permitan dar respuestas y tomar decisiones para realizar transformaciones.

La investigación educativa también se vio atravesada por los avances de la estadística y por el florecimiento de la ciencia de datos y hasta por la evolución del aprendizaje automático. En los años recientes la incorporación de estos procedimientos y estrategias ha delineado un cambio en el escenario cultural de la investigación educativa señalando una tendencia hacia visiones integradoras y comprensivas en los contextos educativos.

En la actualidad, se ha ampliado notablemente la disponibilidad de herramientas digitales, facilitando los procesos de relevamiento y almacenamiento de grandes cantidades de datos provenientes de la administración, de la gestión, de la experimentación o de la observación. Simultáneamente, el acceso a una diversidad de lenguajes, plataformas y software estadísticos favorece un análisis más integrador, preciso y efectivo.

El uso de software ahorra tiempo, dada su flexibilidad en el trabajo con los datos, facilita las tareas rutinarias, agiliza la gestión y el análisis de datos al mismo tiempo que colabora en el tratamiento conjunto de una multiplicidad de variables vinculadas con un mismo proceso. Sin embargo, estas nuevas facilidades disponibles conllevan en sí mismas nuevas dificultades. Es menester destacar que la mera aplicación de herramientas estadísticas o informáticas no es suficiente para una buena comprensión de los fenómenos estudiados, ya que los programas no son capaces de discernir los significados o dimensiones trabajadas. Todo esto otorga un rol protagónico al dominio cabal de la lógica estadística aplicada, a fin de que las conclusiones resulten válidas y las interpretaciones correctas.

La elección del software estadístico está vinculada íntimamente con las necesidades de la investigación, la naturaleza de las variables, con los alcances del estudio, y con la disponibilidad de licencias, así como también, con las habilidades de los usuarios. Entre los softwares más referenciados en todas las áreas del conocimiento y particularmente en educación, podemos mencionar en el área cuantitativa: SPSS, STATA, STATGRAPHICS, Statistica, XLSTAT, SAS, MatLab, MINITAB y funcionalidades de la planilla de cálculo y, la popularización de los CAQDAS (acrónimo en inglés de *Computer Assited Qualitative Data Analysis Software*) de la mano de Atlas.ti, NVivo, MAXqda, Ethngrahp, XSight, HyperRESEARCH, Qualrus y QDA Miner, en el área cualitativa. Algunos de estos paquetes se pueden descargar como versión de prueba, sin embargo, las licencias de los softwares son generalmente onerosas y su valor resulta en ocasiones superior a la disponibilidad de recursos asignados al proyecto de una investigación. Otros rasgos distintivos de algunos de estos programas es su falta de flexibilidad, dado que sus códigos no están disponibles o el acceso se encuentra restringido, de tal forma que no es posible que el usuario los adapte a sus contextos particulares, al tiempo que su interfaz gráfica es poco flexible y resulta limitante en algunas oportunidades.

Además de los mencionados, podemos hacer uso de programas con licencia GNU (acrónimo en inglés de *General Public License*) que es libre y gratuita y permite que los usuarios puedan editar, ejecutar, copiar y/o distribuir el código fuente o software, hasta incluso subirlo a foros o blogs. Entre ellos podemos mencionar el AQUAD, que se utiliza para analizar datos cualitativos, el PSPP para los cuantitativos y lenguajes de programación con enfoque en el análisis estadístico que requieren competencias de programación por parte de los usuarios.

Particularmente, el lenguaje R, diseñado específicamente para el análisis de datos, ha evidenciado un notable desarrollo y actualización de las herramientas de visualización en los últimos años. Es un lenguaje de programación orientado a objetos, distribuido también con licencia GNU, desarrollado por *The R Foundation for Statistical Computing*. Como otros lenguajes, R requiere de un entorno de desarrollo integrado (IDE) como programa de aplicación y, entre otros disponibles, el más utilizado es RStudio. Este programa posee una interfaz compuesta por una consola de comandos que requiere del usuario la escritura de líneas de código y su posterior ejecución. Además, está integrado por un conjunto de librerías, sistema de ayuda e instrucciones que posibilitan explorar, realizar cálculos, modelar y visualizar datos para extraer significado de información. Este entorno de trabajo está disponible en distintas plataformas (Microsoft Windows, Linux o Macintosh), es gratuito y de libre acceso. Invita a sus usuarios a compartir y distribuir sus desarrollos, promueve el trabajo colaborativo entre sus usuarios a partir de la difusión y mejora de los códigos ya elaborados que son compartidos generalmente en RPubs (s.f).

En este trabajo hemos focalizado el interés en la aplicación de métodos de análisis estadístico multivariado o multivariante (AM), que brindan una visión integral y contiene a los análisis univariados e incluso bivariados como enfoques reducidos del mismo. Recientemente se ha evidenciado el crecimiento de su aplicación en diferentes áreas del conocimiento tales como economía, biología, sociología y medicina. La educación también ha aprovechado la contribución de estos métodos tanto para inferir y deducir patrones y estructuras subyacentes, como para contrastar hipótesis y supuestos en forma eficiente y veloz. Asimismo, estos han facilitado la construcción y validación de modelos que permiten la interpretación de los complejos fenómenos de interés.

El AM comprende un conjunto de conocimientos, técnicas, estrategias y metodologías que permiten la interpretación, análisis y descripción simultánea de varias características o atributos (variables) sobre un conjunto de individuos (objetos o unidades de análisis) correspondientes a una misma variedad designada como población de interés. Este conjunto de técnicas comporta una visión multidimensional de la realidad explorada, puesto que facilita el tratamiento, la visualización y la interpretación de grandes bases de datos, tanto respecto de la cantidad de observaciones como de la cantidad de variables. Además, extiende el alcance de la inferencia estadística, incluyendo a los análisis univariado y bivariados como casos particulares.

En los últimos años, dada la expansión de la ciencia de datos y del aprendizaje automático, se ha difundido en forma destacable la aplicación de técnicas o métodos del AM. Esto ha generado nuevas demandas de software y ha favorecido el desarrollo de algoritmos específicos y software para su ejecución. En términos generales, en el campo de las variables se puede establecer una clasificación vinculada a la existencia de una relación de dependencia o de interdependencia entre las mismas. En algunos casos que se explora una relación de dependencia mediante un modelo relacional, donde una o más variables respuesta se explican a partir de una o más variables predictoras, cuantitativas o cualitativas, como el caso de los modelos de regresión lineales o no lineales, simples o multivariados. Dentro del conjunto de estos modelos, en los cuales la variable dependiente es única, podemos mencionar el análisis de la varianza (ANOVA), la regresión logística, la regresión de Cox, la regresión de Poisson, los modelos no lineales con efectos fijos, aleatorios o mixtos, el análisis discriminante y el análisis factorial, entre otros. En cambio, si las variables respuesta son dos o más podemos mencionar el análisis de estructuras de covarianza, modelos de ecuaciones estructurales, el análisis de la varianza multivariado (MANOVA) y análisis de covarianza. Por otro lado, si estamos en presencia de una relación de interdependencia, sin distinción entre variables explicativas y respuesta, se ubica análisis log-lineal, análisis de componentes principales (ACP), análisis de conglomerados o clusters y el análisis de correspondencias múltiples (ACM). (Chan, Badano y Rey, 2020)

A partir de lo expuesto, la principal motivación de este trabajo es explorar e identificar información relevante en investigaciones educativas de nivel superior, con incidencia en aspectos académicos o pedagógico-didáctico, que hayan aprovechado los beneficios del AM y del lenguaje R como oportunidad para ampliar el horizonte y los alcances de los estudios. Asimismo, interesa describir las técnicas utilizadas con mayor frecuencia en estos estudios y analizar la evolución de la incorporación de éstas en este ámbito del

conocimiento. Por último, se pretende explorar la contribución diferencial al horizonte de la investigación producido como resultado de la aplicación de estos recursos.

Metodología

La metodología utilizada para la revisión bibliográfica es la sugerida por Okoli (2015) la cual se organiza mediante una secuencia de procedimientos que listamos a continuación: i) identificación del objetivo de la revisión, ii) elaboración del protocolo, iii) determinación de criterios de exclusión e inclusión de artículos, iv) búsqueda de la literatura, v) extracción de datos tanto cualitativos como cuantitativos, vi) evaluación de calidad de los resultados, vii) síntesis del relevamiento y, viii) elaboración del informe de revisión.

Este trabajo de revisión sistemática ha examinado la literatura existente y disponible de trabajos de investigación nacionales e internacionales, desarrollados en el área educativa del nivel superior, vinculados con aspectos académicos o pedagógico-didáctico, durante el período 2018-2019. La finalidad es describir las características que tienen los últimos trabajos de investigación educativa que aprovechan el lenguaje R como herramienta y que aplican estrategias de AM, así como distinguir el valor agregado que pudiera aportar el aprovechamiento de estos recursos. Particularmente, se ha focalizado en este ámbito, puesto que se busca destacar sus potenciales aportes en una ciencia social en la cual su aplicación es incipiente.

El protocolo, a través que se condujo el proceso de revisión, estuvo orientado por las siguientes preguntas:

1. ¿Qué cantidad de las investigaciones del área educativa que aplican metodologías propias del AM, correspondientes al nivel superior, se han publicado en el período 2018-2019 y han referenciado en sus metodologías el uso del lenguaje de programación R?
2. ¿Cuáles son las principales categorías temáticas en las que se pueden clasificar los documentos revisados?
3. ¿Cuál es el alcance de las investigaciones presentadas en los documentos revisados?
4. ¿Cuáles son las técnicas de AM más aplicadas en los documentos revisados?
5. ¿Qué visualizaciones son las más frecuentes en estas publicaciones?
6. ¿De qué manera las técnicas de AM y el lenguaje R contribuyen en las investigaciones?

En una etapa preliminar de la búsqueda, se definieron tres palabras claves relacionadas con el tema de interés: análisis multivariado, educación superior y lenguaje R. Cabe destacar que el lenguaje R es referenciado con sinónimos, por ello se incluyeron términos tales como *entorno R, software R, Rstudio, software Rstudio, R Project y software R Proyect*, situación análoga con educación superior, en el que se consideraron *nivel superior, educación universitaria* y *educación superior no universitaria*. De esta misma forma se ha procedido con *análisis multivariado* y *análisis multivariante*. Estos términos de búsqueda fueron explorados tanto en las secciones de títulos, en palabras claves, resúmenes y en el cuerpo del texto, conectados mediante los operadores booleanos AND y OR para seleccionar publicaciones en el período 2018-2019, con el propósito de obtener una mirada actualizada en la revisión.

Los criterios de inclusión han involucrado publicaciones que contengan las palabras clave destacadas, realizadas en idioma español, pudiendo incorporarse trabajos de tesis, revistas indexadas y con referato a fin de garantizar la evaluación y control de calidad realizada. Se excluyeron trabajos correspondientes a otros niveles educativos y áreas del conocimiento, los que no tenían incidencia en lo académico-pedagógico, los duplicados en distintos repositorios y los que utilizaban otros recursos de software.

Para la búsqueda de información se utilizaron como motores de búsqueda *Google Scholar* y *Semantic Scholar*; colecciones de revistas, libros y recursos de investigación tales como *Redalyc, SciELO y Wiley on Line Library*, además de bases de datos especializadas educación como *ERIC* y *Redined*, las que arrojaron 168 documentos (al 17.01.2020). La selección de estudios vinculados al objetivo se ha realizado mediante una inspección

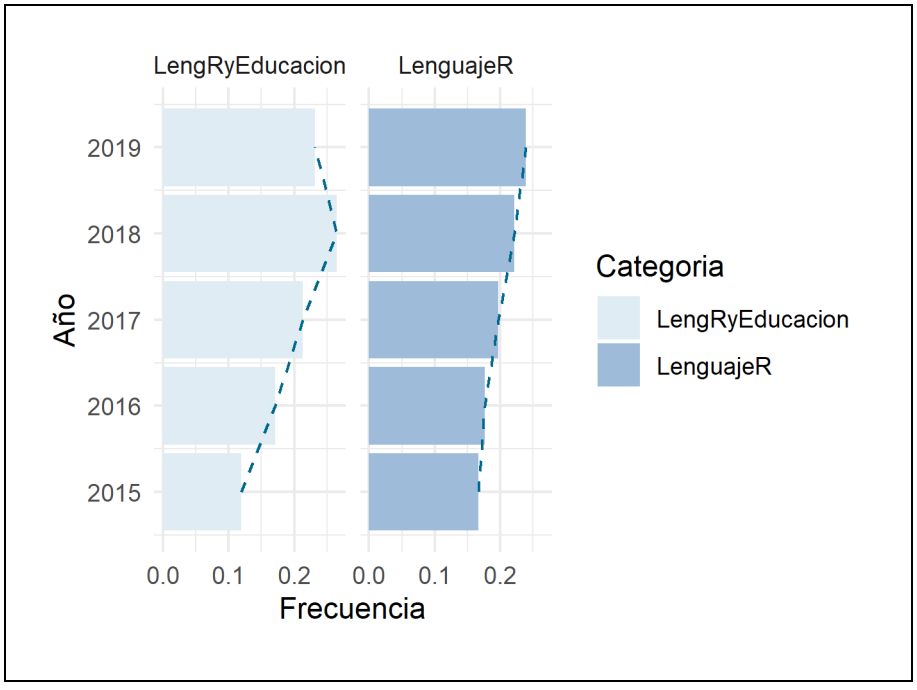
manual, aplicando los criterios de selección descriptos. Finalmente, solamente 10 de los documentos fueron seleccionados como material definitivo de análisis para el presente trabajo.

A continuación, se procedió a una clasificación minuciosa y una reorganización de la literatura seleccionada durante el proceso de búsqueda. Con este propósito se confeccionó una matriz que considera para cada uno de los documentos los siguientes campos: link del documento, tipo de fuente, título, autores, año de publicación, tipo de trabajo (artículo de investigación, reseña de tesis, tesis, etc.), los objetivos, las palabras claves, los enfoques y los resultados. Además, se registró en la misma matriz, el tipo de técnica o método específico del AM, como también las pruebas estadísticas y gráficos utilizados, acompañados de una sección de observaciones. A partir de la síntesis de información recogida de los documentos se procedió a realizar el análisis de la misma con el objetivo de dar respuesta a las preguntas de investigación planteadas.

RESULTADOS

En una etapa previa a la búsqueda de la cadena de términos claves, nos ha parecido pertinente explorar en los motores de búsqueda, base de datos y colecciones mencionadas, la frecuencia de aparición del término *Lenguaje R* (así como de sus sinónimos) en documentos durante el período 2015-2019. Además, con el propósito de refinar la búsqueda incluimos la cadena *Lenguaje R* (y sus sinónimos) AND *Educación*. En el primer caso se han obtenido 2.854.118 resultados y en el segundo 97.983.

Ilustración 1: Cantidad de publicaciones período 2015-2019



Fuente: Elaboración propia a partir de datos de Google Scholar, Semantic Scholar, Redalyc, Scielo, Wiley on Line Library, ERIC y Redined (17.01.2020)

Como se puede apreciar en la ilustración 1, la frecuencia del término *Lenguaje R* en las distintas publicaciones ha evidenciado un crecimiento sostenido durante los últimos años, con un pico en publicaciones vinculadas con educación en el año 2018, de lo que se infiere una ganancia de territorio por parte del software en la comunidad académica en general y en la educativa en particular. Esta evidencia ya fue manifestada por Elosua (2009) en su obra "¿Existe vida más allá del SPSS? Descubre R", quien hace más de una década ya compartía con la comunidad académica las propiedades de la herramienta para el tratamiento y análisis de datos aplicado tanto a la docencia como a la investigación. Esta leve reducción en la cantidad de publicaciones en educación

que han trabajado con R durante el año 2019, podría tener como explicación que los entornos que albergan a R entre sus lenguajes, como Infostat, Python o Julia, pudieron haber captado algunas de las investigaciones en general y en educación en particular. Esta observación está alineada con la sexta edición de la encuesta llevada a cabo por Burtch Works LLC (2019), en la cual se destaca que los profesionales en análisis de datos y los estudiantes de nivel superior, durante el último semestre 2018 y el primero del 2019 prefirieron utilizar en sus trabajos Python respecto de otras herramientas como R y SAS. Y específicamente en el área de ciencias sociales se mostró una muy leve tendencia de R por sobre Phyton.

Para enfocarnos en nuestro trabajo específico, delimitamos secuencialmente la búsqueda, como se expuso anteriormente, con la cadena (“análisis multivariado” OR “análisis multivariante”) AND ("educación superior" OR “nivel superior” OR “educación universitaria” OR “educación superior no universitaria”) AND ("Lenguaje R" OR “Software R” OR "Rstudio" OR ”R Project” OR “entorno R” OR “software Rstudio” OR “software R Project”) y se han obtenido 168 documentos, de los cuales 59 son en español y 10 se ajustaron a los criterios de selección, distribuidos como puede apreciarse en la tabla 1.

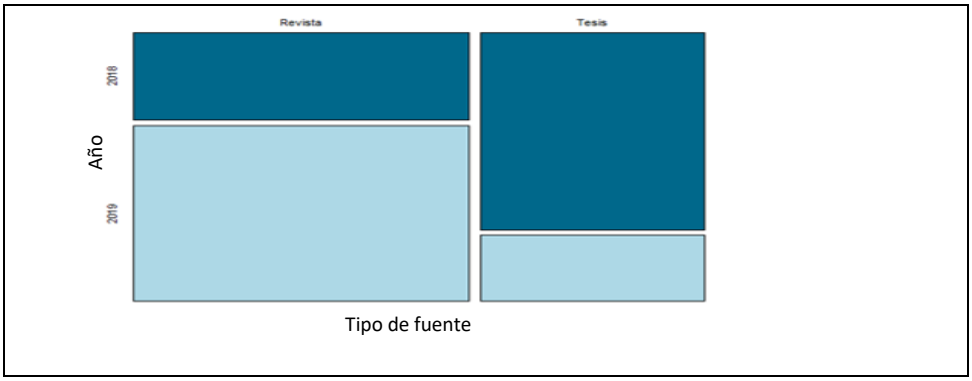
Tabla 1: Detalle de selección de fuentes

Total de documentos encontrados	Total de documentos en español		Documentos seleccionados con foco de esta investigación			
			Año	Cantidad	Tipo de fuente	Tipo de trabajo
168	59		2018	2	Artículo de revista	Investigación aplicada
				3	Tesis	Investigación
	2018	2019	2019	4	Artículo de revista	Investigación aplicada
	39	20		1	Repositorio de Tesis	Investigación

Fuente: Elaboración propia

Aunque el 66% del total de los documentos encontrados en español corresponden al año 2018, se han seleccionado cantidades similares para el análisis en profundidad. En estos se aprecia, en el año 2019, un crecimiento en el uso de las técnicas multivariadas con el lenguaje R en los trabajos de investigación aplicada en el área de educación superior, lo cual puede observarse en el gráfico de mosaicos de la Ilustración 2.

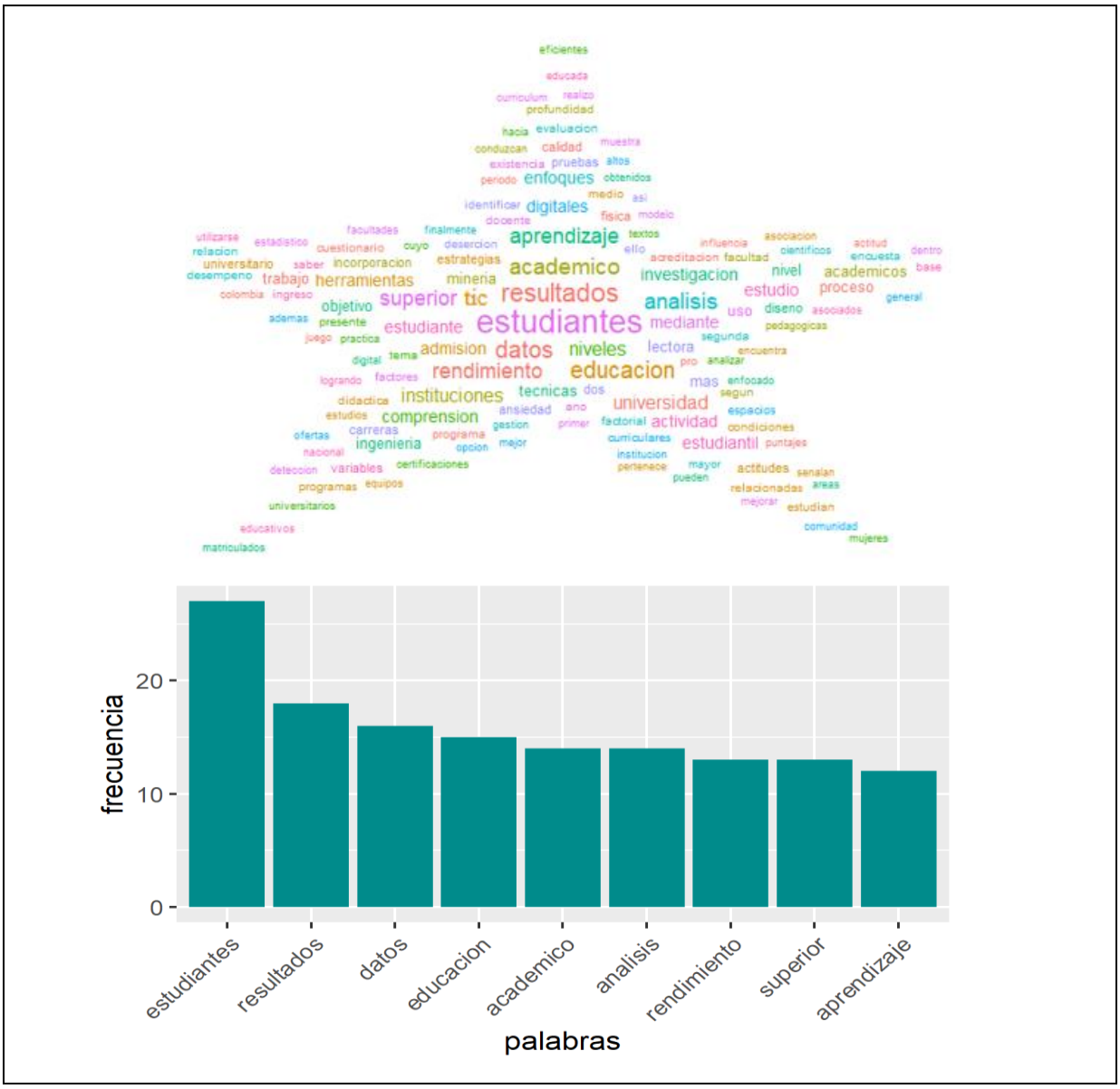
Ilustración 2: Distribución de documentos por año y tipo de fuente



Fuente: Elaboración propia

Con el propósito de analizar las temáticas presentes en los documentos explorados, se presenta un análisis de contenido bajo la categoría “análisis temático” a partir de la frecuencia de aparición de términos en sus palabras claves y resúmenes.

Ilustración 3: Análisis temático



Fuente: Elaboración propia

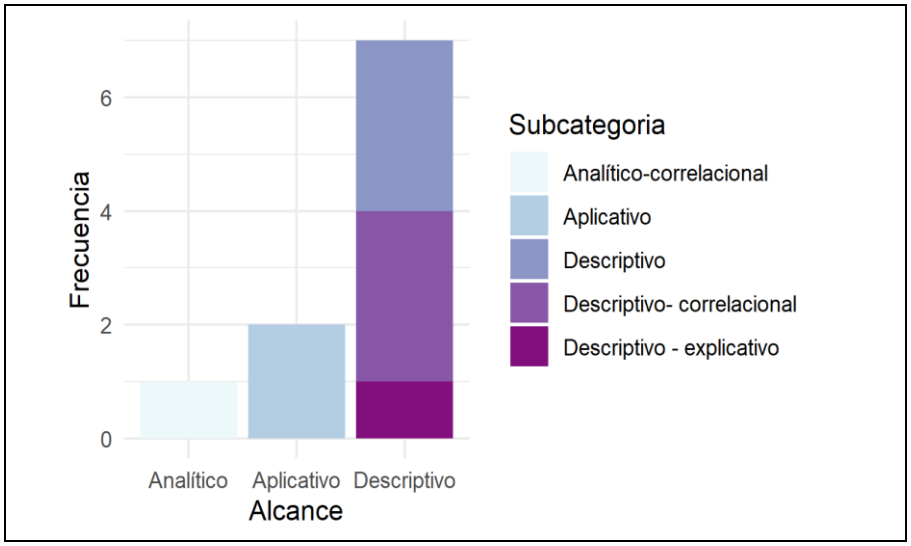
En la ilustración 3, se representan las densidades de palabras que referencian al análisis temático. En la nube de palabras se puede apreciar la variación de tamaño y color de las palabras, donde estos atributos son proporcionales a la frecuencia de cada término.

Particularmente en el diagrama de barras, se observa que los términos utilizados con mayor frecuencia son “estudiantes”, “resultados” (referidos a resultados de los análisis), “datos”, “educación”, “académico”, “análisis”, “rendimiento”, “superior” y “aprendizaje”, de los cuales podemos concatenar “análisis de datos” y “educación superior”. Vale aclarar que, más allá que la frecuencia de cada palabra es sencilla de obtener con la medicación de un software, estos no consideran los posibles cambios de significación según el contexto. Tal es el caso del término “resultados”, que se ubica en el segundo lugar debido a que el motor de búsqueda lo detecta a raíz de que los autores, dentro de sus resúmenes y palabras claves, lo referencian sistemáticamente previo a la narración de su síntesis del trabajo.

Este análisis de contenido efectuado con R coincide con los resultados obtenidos a partir de la reducción de datos mediante codificaciones y agrupamientos realizados de forma manual donde se evidencia que las líneas de investigación donde se aplican técnicas AM con el apoyo del lenguaje R están enfocadas en temas específicos. Entre estos podemos destacar el rendimiento académico de los estudiantes, el uso y/o integración de las TIC, los enfoques y procesos de aprendizaje y las estrategias de admisión, retención, deserción y permanencia de los estudiantes en este nivel educativo.

En cuanto al alcance de las investigaciones en los documentos analizados, el 70% se corresponde con un enfoque descriptivo, que persiguen especificar factores, características o rasgos de los fenómenos analizados. El resto de los trabajos son del tipo aplicado y analítico, los que identifican problemáticas sobre los que se interviene. En la ilustración 4 se presenta un desglose de las características del alcance de estos trabajos de investigación.

Ilustración 4: Tipos de alcance de las investigaciones analizadas



Fuente: Elaboración propia

Uno de los problemas fundamentales del AM tratado desde diversas perspectivas, es la reducción de la dimensionalidad. En este contexto para datos cuantitativos la técnica más aplicada es el ACP, mientras que para datos cualitativos son el análisis factorial, las tablas de contingencia, el test de Chi cuadrado de Pearson y el análisis de correspondencias simples o múltiples (AC).

No es sorprendente pues que entre las técnicas del AM más aplicadas en los trabajos explorados hallemos el ACP (Almandoz et.al, 2019; Autor, 2018; González, 2019; Holgado Apaza, 2018; Méndez, Tovio, y Vertel, 2018), pero también aparece el análisis de conglomerado o cluster (Autor, 2018; González, 2019; González-García, Sánchez-García, Nieto-Librero, y Galindo-Villardón, 2019; Morales, 2018). Asimismo, técnicas de análisis factorial múltiple (Méndez, Tovio, y Vertel, 2018) y de regresión logística (Méndez, Tovio, y Vertel, 2018; Tapasco-Alzate, Ruiz-Ortega, Osorio-García y Ramírez-Ramírez, 2019). Por otro lado, es interesante destacar que en dos de los trabajos analizados aparece una combinación del aprendizaje automático, específicamente árboles de regresión y clasificación (random forest) (González, 2019; Holgado Apaza, 2018) y de análisis factorial confirmatorio y redes neuronales (Jácome-Ortega, A. et. al, 2019), lo cual podría ser el origen de una nueva ampliación del horizonte y del alcance de los estudios.

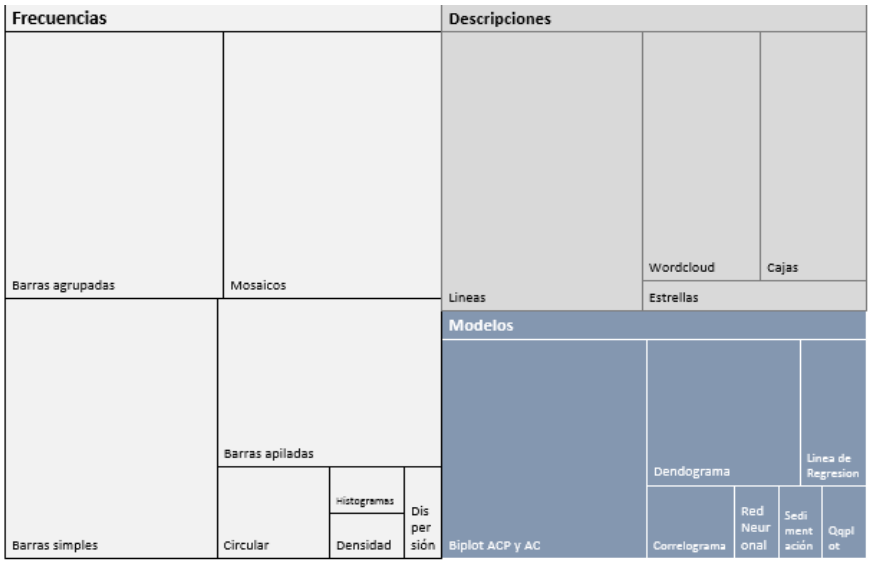
Dado que surge del relevamiento de datos que la técnica ACP es una de las más aplicadas, consideramos pertinente destacar algunos de sus aspectos característicos. Se denominan componentes principales a variables construidas a partir de las originales como combinación lineal de las mismas, pero de forma tal que capten la máxima información disponible en la base de datos y al mismo tiempo que resulten independientes entre sí.

Estadísticamente esto significa que no hay información redundante, es decir que lo que dice la primera componente no se repite en ninguna de las siguientes. Estas nuevas variables o componentes, en general aluden a variables o características latentes, que posibilitan descubrir y describir patrones de estructura a partir del aprovechamiento de la correlación o asociación entre las variables originales. Este análisis es particularmente enriquecedor cuando existe una alta correlación entre las variables originales, mientras que si las variables disponibles son no correlacionadas este procedimiento carece de sentido. Cabe aclarar que las componentes pueden ser de forma o de tamaño. En el primer caso contrastan los valores de las variables originales mientras que en el segundo realizan un promedio ponderado de ellas. En general podemos decir que se trata de una técnica descriptiva y no tiene supuestos, lo que posibilita su aplicación en cualquier caso y en muchas oportunidades se combina con otras técnicas del AM como el análisis de clusters o el análisis discriminante.

Respecto de las gráficas utilizadas para ofrecer una perspectiva visual de los resultados de los análisis hemos definido tres categorías: las que permiten visualizar frecuencias, las que describen fenómenos o vinculaciones entre ellos y las que se vinculan con modelos específicos. La visualización ha cobrado una importancia fundamental en el contexto de la ciencia de datos. El lenguaje R, en este sentido, ha evolucionado notablemente en sus paquetes o bibliotecas, donde utilizando los avances del entorno ggplot2 se han mejorado muchas librerías. Listar la diversidad de gráficos hoy disponibles no es una tarea sencilla dado que en cada tipo de análisis se han incorporado varias herramientas en los años recientes y este proceso es de crecimiento continuo. Además, cabe destacar que muchos de los avances están documentados y disponibles para los usuarios en el sitio web colaborativo RPubs donde también se pueden encontrar manuales y tutoriales explicativos con contenidos teóricos y múltiples ejemplos.

Como se aprecia en la ilustración 5, los gráficos utilizados con mayor frecuencia en los documentos analizados son los diagramas circulares, los gráficos de mosaicos, los gráficos de líneas, las representaciones con barras agrupadas, simples, piramidales, paralelas y los biplots correspondientes a los análisis de reducción de dimensionalidad.

Ilustración 5: Gráficas más utilizadas



Fuente: Elaboración propia

Cabe preguntamos ahora si la mera aplicación de técnicas multivariadas es suficiente para conducirnos a una mejora en el análisis y tratamiento de los datos. Para poder evaluar esto abordamos un esquema comparativo. Para ello se seleccionaron dos pares de documentos donde la temática de cada par es muy similar, pero en uno de los documentos se han aplicado análisis univariados mientras que en el otro se ha aplicado AM.

El primero de los pares está integrado por las tesis “Análisis de planificación en el uso de las Tecnología de las Información y Comunicación (TIC) en los cursos virtuales de pregrado en la PUCP basado en la MATRIZ TIC de Planificación” (Alfaro Salas, 2017) y “Mecanismos de gestión para incorporar herramientas digitales en los espacios curriculares de instituciones de nivel superior” (Autor, 2018), los que relevan parte de sus datos con un instrumento adaptado de la “Matriz de planificación TIC”, propuesto por Lugo y Kelly (2011), la cual permite perfilar un estado de situación en relación a la incorporación de las TIC en las instituciones.

El primero de estos dos estudios trabaja con 6 dimensiones o conjuntos de variables: gestión y planificación, las TIC y el desarrollo curricular, desarrollo profesional docente, cultura digital, recursos e infraestructura e institución. El autor utiliza un enfoque univariado, describiendo el comportamiento de las características de la muestra estudiada y detallando medidas de tendencia central sin vincular las dimensiones entre sí e incorporando tablas y gráficos circulares para la visualización de los datos. Mientras que, en el segundo de los estudios, se realiza una observación univariada, una mirada bivariada y se converge hacia un AM que las integra. Para el análisis univariado, se utilizan como en el primer trabajo, medidas de tendencia central, acompañadas por tablas y gráficos circulares y de barras; en el análisis bivariado, se aplican visualizaciones con correlogramas y pruebas estadísticas para analizar y cuantificar la existencia de asociación entre dos variables. Y, por último, en el AM se usan las técnicas de AC y ACP ilustrados con boxplots para el análisis de distribución de una variable entre grupos y de biplots con elipses de concentración para una mejor visualización de los datos en forma integrada, además de análisis de conglomerados orientados por las exploraciones univariadas realizadas previamente. En este segundo estudio, la autora trabaja con dimensiones similares: gestión y planificación, herramientas digitales y desarrollo curricular, desarrollo profesional docente, cultura, recursos e infraestructura e institución y comunidad. Sin embargo, establece la correlación entre las dimensiones, se estudia la posición de los individuos por institución en función de las dimensiones, se compara a las instituciones en función del posicionamiento en las componentes principales pudiendo definir dos nuevas variables que generaron una partición de las originales y permitieron delinear las tipologías diferenciales de los individuos por institución y explicar las diferencias entre aquellas. Una vez lograda la definición de las tipologías es posible analizar otras instituciones que inicialmente no fueron incluidas en el estudio en el marco de un análisis confirmatorio.

En consecuencia, mientras el primer estudio se centró en la descripción de cada dimensión de análisis por separado, el segundo permitió analizar cómo funcionan conjuntamente las distintas dimensiones y en qué medida explican, desde la percepción de los encuestados, el estado de situación de la integración de las herramientas digitales en los espacios curriculares de distintas instituciones.

En la misma línea encontramos diferencias entre las investigaciones “Comprensión lectora y rendimiento académico en estudiantes de educación superior” (Martínez, Paredes, Rosero y Menjura, 2013) y “Niveles de comprensión lectora de textos científicos en estudiantes de ingeniería” (Almandoz et al., 2019) los cuales persiguen la medición de la comprensión lectora en estudiantes de nivel superior.

El primero de los dos artículos es empírico analítico correlacional y de corte transversal. Analiza el comportamiento de dos conjuntos de estudiantes: el primero correspondiente a la carrera de ingeniería y el otro de la carrera de psicología, constituyendo una muestra total de 60 estudiantes. Para el relevamiento de datos acerca del nivel de comprensión lectora que poseen los mencionados, se construye un instrumento subdividido en distintos niveles. Para la etapa de análisis se establecen correlaciones usando el coeficiente y test de Spearman y Pearson, se construyen intervalos de confianza para el rendimiento por carrera y distingue entre población rural y urbana. En el análisis cuantitativo se utilizan medidas de tendencia central y se correlacionan medidas de rendimiento académico; edad-rendimiento; comprensión lectora con la escala de acceso y recuperación; escala de integración e interpretación; escala de reflexión y evaluación; comprensión lectora-lugar de pertenencia y particularmente se realizan correlaciones entre cada una de las escalas de comprensión lectora y la variable rendimiento académico. El segundo estudio trabaja con 370 estudiantes de todas las carreras de Ingeniería de la Regional Buenos Aires de la Universidad Tecnológica Nacional que se inscribieron en Inglés Técnico I que, puesto que no tiene correlatividades, pueden corresponder a cualquier año de la carrera. Las autoras trabajan con un cuestionario de comprensión lectora elaborado específicamente para este

estudio que consta de 10 preguntas, de las cuales 7 son de selección múltiple y 3 son abiertas. Las preguntas evalúan la comprensión proposicional, en sus aspectos tanto micro como macroestructurales, y la comprensión de factores enunciativos claves para la comprensión de textos académicos. En este caso las autoras correlacionan junto con el rendimiento académico, la disponibilidad horaria para el estudio de los participantes medida a través de la existencia de o no de un trabajo formal, el tipo de gestión de la escuela media cursada y la experiencia previa en comprensión lectora realizada o no en la educación secundaria. Para ello, se construyeron variables como nota y se aplicó un ACM, ingresando como variables originales a las recientemente mencionadas. A través de este análisis se pudo construir una segmentación de los estudiantes en tres grupos y describir las características distintivas de cada uno de estos segmentos. También se presenta en el artículo un ACP por grupos definidos por las variables categóricas mencionadas y la variable nota construida. De esta manera se logró identificar características específicas de los estudiantes en función tanto de las variables categóricas como de las variables continuas. El resultado de estos análisis condujo a la definición de dos nuevas variables, una de ellas vinculada con el grado de avance del estudiante en su formación, es decir, con la edad, cantidad de materias cursadas, cantidad de materias aprobadas/pendientes y otra relacionada con la rapidez de avance de los participantes en el trayecto de su formación académica. Asimismo, este estudio utilizó diversas representaciones gráficas entre las cuales podemos destacar los diagramas circulares, los gráficos de barras simples, el boxplot o diagrama de caja y el biplot.

En ambas comparaciones se evidencia que los enfoques son bien distintos y esto impacta natural e inmediatamente en el alcance y horizonte de las conclusiones que se derivan de los mismos. Quedan delineadas las potencialidades del AM respecto al univariado y bivariado, ya que la exploración multivariada favorece la construcción de un enfoque integral de las situaciones en el contexto de variables educativas tanto cualitativas como cuantitativas, permitiendo en algunos casos integrar ambas en un único análisis. Por otro lado, el método de ACP, ha posibilitado la exploración de bases de datos complejas y la construcción de patrones y estructuras de relación, así como de vinculación de variables. En los dos segundos estudios mencionados se referencia el uso del lenguaje de programación R, aspecto que ha favorecido en la variedad de gráficas presentadas a partir de las vastas librerías que ofrece el IDE.

Conclusiones

En la actualidad, la mayoría de los problemas abordados en investigación educativa del nivel superior requieren la integración de múltiples dimensiones, algunas de las cuales son observables y han sido registradas y otras son latentes, las que surgen del análisis estadístico y se definen a partir de las originales. La metodología del AM facilita la construcción y el descubrimiento de unas pocas dimensiones que permiten caracterizar en forma sencilla a la complejidad de información inicialmente disponible. Es decir, nos permite comprender más profundamente el fenómeno de interés a través del tipo de relaciones que pueden establecerse entre las variables. Por otro lado, la disponibilidad de herramientas de software libre, como es el caso del lenguaje de programación R, nos permite manejar grandes volúmenes de datos, manipularlos, visualizar su comportamiento, describir sus estructuras, resumir las redundancias y detectar la presencia de datos anómalos integrándolos al análisis mediante la aplicación de metodologías robustas.

Uno de los problemas actuales con los que nos enfrentamos al momento de poner en marcha una investigación, es que seguramente dispondremos de grandes cantidades de datos, los que requirieren de la realización de inferencias de valor a partir de los mismos. En esta línea, han surgido lenguajes de programación focalizados en el tratamiento estadístico, como es el caso de R. R surgió con el propósito de mejorar las prestaciones del software libre del lenguaje S, creado en 1995, por Ihaka y Gentleman, siendo utilizado en la actualidad en los campos de la ciencia de datos, minería de datos, aprendizaje automático y en distintas disciplinas científicas.

En esta investigación centramos el objetivo en la exploración especifica de las potencialidades del uso del Lenguaje R en cuanto a las oportunidades que este recurso de vanguardia ofrece dentro del campo de la investigación educativa del nivel superior, específicamente en relación con el aprovechamiento del análisis simultáneo de varias variables. Se ha focalizado en él por ser pionero en el campo de la estadística de uso libre

y gratuito que ha formado una comunidad que potencia sus alcances mediante bibliotecas muy actualizadas, foros de discusión, reuniones científicas y colaboración didáctica disponible para los usuarios de tutoriales, ejemplos y desarrollos de línea de códigos. A pesar de ello, su acceso en las ciencias sociales ha mostrado un ritmo más lento dado que requiere de competencias en programación que no son frecuentes en los profesionales de esta área. Sin embargo, cursos de capacitación para el uso de este lenguaje en la investigación social son ofrecidos por múltiples universidades y centros privados de formación.

Surge de la comparación de estudios con temas similares abordados en forma univariada o bivariada frente a estudios multivariados una importante ventaja de estos últimos frente a los primeros en su capacidad para integrar y profundizar múltiples perspectivas. Por otro lado, se aprecia un aumento progresivo en el uso del Lenguaje R en general y en particular en ciencias sociales. En este sentido, el crecimiento ha sido sostenido y las investigaciones en educación superior no permanecieron ajenas a esta evolución. Sin embargo, un leve descenso en la cantidad de publicaciones se observa en 2019 debido a que la comunidad científica dispone de otras oportunidades como es el caso de Phyton para los mismos objetivos. Es decir, a pesar de la calidad de competencia de este lenguaje frente a otros softwares privativos del mercado, la evolución de otros lenguajes de programación y entornos Phyton y Julia, puede haber incidido en un descenso en su referenciación.

De este modo, se han planteado dos nuevos desafíos en la actualidad para los profesionales del área de ciencias sociales en general y de la educación superior en particular. Por un lado, el desarrollo de habilidades de programación elemental para acceder al uso y aprovechamiento de estos lenguajes y, por otro, la ampliación del horizonte de las miradas utilizando las técnicas multivariadas. Este avance resulta imperativo frente a la complejidad del escenario del conocimiento en la actualidad.

Referencias bibliográficas

Alfaro Salas, E. (2017). *Análisis de planificación en el uso de las Tecnología de las Información y Comunicación (TIC) en los cursos virtuales de pregrado en la PUCP basado en la MATRIZ TIC de Planificación* [tesis de maestría*]*. Perú: Pontificia Universidad Católica del Perú.

Almandoz, P. et al. (2019). Niveles de comprensión lectora de textos científicos en estudiantes de ingeniería. *Revista Argentina de Educación Superior*, 78-95.

Bisquerra Alzina, R (coord). (2009). *Metodología de la investigación educativa.* Madrid, España: La Muralla.

Burtch Works LLC. (2019, 21 de agosto). *2019 SAS, R o Python 2019 Survey Update: Which Tool do Data Scientists & Analytics Pros Prefer?* Recuperado de: https://www.burtchworks.com/2019/08/21/2019-sas-r-or-python-survey-update-which-tool-do-data-scientists-analytics-pros-prefer/

Chan D., Badano, C. y Rey, A. (2020). *Análisis Inteligente de Datos con R: con aplicaciones a imágenes*. Buenos Aires, Argentina: Edutecne.

Elosua, P. (2009). ¿Existe vida más allá del SPSS? Descubre R. *Psicothema, 21*(4), 652-655.

Galli, M. (2018). *Mecanismos de Gestión para incorporar herramientas digitales en los espacios curriculares de Educación Superior* [Tesis en prensa]. Buenos Aires: UNTREF.

González, C. (2019). *Análisis por minería de datos del impacto de los sistemas de calidad de las instituciones de educación superior en los resultados de las pruebas saber pro enfocado a los programas de ingeniería industrial.* [Tesis]. Colombia: Universidad Tecnológica De Bolívar.

González-García, N., Sánchez-García, A., Nieto-Librero, A. y Galindo-Villardón, M. (2019). Actitud y enfoques de aprendizaje en el estudio de la Didáctica General. Una visión multivariante. *Revista de Psicodidáctica*, 24(2), 154-162.

Holgado Apaza, L. (2018). *Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios 2018.* [Tesis]. Perú: Universidad del Altiplano.

Jácome-Ortega, A. et. al (2019). Análisis temporal y pronóstico del uso de las TIC, a partir del instrumento de evaluación docente de una Institución de Educación Superior. *Revista Ibérica de Sistemas e Tecnologías de Informação*, 22(8), 399–412

Lugo, M. y Kelly, V. (2011). *La matriz TIC. Una herramienta para planificar las Tecnologías de la Información y Comunicación en las instituciones educativas.* Buenos Aires, Argentina: IIPE-Unesco Buenos Aires.

Martínez, A., Paredes, L., Rosero, S. y Menjura, M. (2015). *Comprensión Lectora Y Rendimiento Académico En Estudiantes De Educación Superior.*

Méndez, M., Tovio, J. y Vertel, M. (2018). Análisis multivariado de los factores socio-económicos asociados al rendimiento en las pruebas saber pro 2016: el caso de los estudiantes de licenciatura en matemáticas en Colombia. En Valbuena, S., Vargas, L. y Berrío, J. (Eds.), *Encuentro de Investigación en Educación Matemática* (pp. 82-88). Puerto Colombia, Colombia: Universidad del Atlántico.

Morales, N. (2018). *Aplicación de la minería de datos a los registros académicos de los estudiantes de la Universidad Nacional Santiago Antúnez de Mayolo – Huaraz, periodo 2000-2015* [Tesis]. Perú: Universidad Nacional Santiago Antúnez De Mayolo.

Okoli, C. (noviembre de 2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems, 37*(43), 879-919.

Tapasco-Alzate, O., Ruiz-Ortega, F., Osorio-García, D. y Ramírez-Ramírez, D. (2019). Deserción estudiantil: incidencia de factores institucionales relacionados con los procesos de admisión. *Educación y Educadores*, 22 (1), 81-100.

Fuentes

Google Scholar https://scholar.google.com/

Educational Resources Information Center. https://eric.ed.gov/

Redalyc - *Sistema de información científica Redalyc*. https://redalyc.org/

Redined -*Red de información educativa*. *https://redined.mecd.gob.es/xmlui/*

RPubs (s.f). *RPubs by Rtudio*. https://rpubs.com/

SciELO - *Scientific electronic library online.* https://scielo.org/es/

Semantic Scholar https://www.semanticscholar.org/

Wiley on line Library https://onlinelibrary.wiley.com/

Fecha de presentación: 28/2/2020

Fecha de aprobación: 7/5/2020