Lo que todo investigador educativo cubano debiera conocer: el entorno informático R

What every Cuban educational researcher should know: the computer environment R

Ensayo

Paul A. Torres Fernández[1]

paul@rimed.cu / paulantoniotorresfernandez@gmail.com

Resumen

A partir del hecho de que el entorno estadístico **R** se ha convertido en una herramienta informática de código abierto de uso muy extendido, y de que existen puntos clave del proceder investigativo en el campo educacional cubano que deben ser perfeccionados, en el presente trabajo se argumentará hasta qué punto esos *nodos esenciales* del proceder investigativo requieren ser mejorados y, a la par, se fundamentará en qué medida **R** constituye una herramienta eficaz para conducir dicho perfeccionamiento.

Palabras clave: R, RStudio, entorno estadístico, investigación educativa, Cuba.

Abstract

Starting from the fact that the statistical environment **R** has become an open source computer tool of widespread use, and that there are key points of the investigation in the Cuban educational field that must be perfected, in the present work it will be argued to what extent these *essential nodes* of the research procedure need to be improved and, at the same time, will be prove on what measure **R** constitutes an effective tool to conduct this improvement.

Keywords: R, RStudio, statistical environment, educational research, Cuba.

[1] Licenciado en Educación especialidad Matemática. Doctor en Ciencias Pedagógicas y Doctor en Ciencias. Investigador Titular del Instituto Central de Ciencias Pedagógicas, Cuba. http://orcid.org/0000-0002-7862-2737

Introducción

El entorno informático **R** (mejor **RStudio**, un software más *amigable*, dependiente de él), se ha convertido en una *revelación* en la investigación científica de numerosos campos del saber, que incluyen a las ciencias sociales y humanísticas, y dentro de ellas a las *ciencias de la educación*.

R constituye un conjunto de **librerías** e instrucciones informáticas que posibilitan, con su actuación integrada, realizar cálculos, gráficos, pruebas y modelaciones estadísticas, dando cobertura prácticamente a todas las técnicas y métodos estadísticos desarrollados hasta el presente.

Algunas de esas técnicas estadísticas de avanzada no están disponibles siquiera en paquetes corporativos de renombre, como **SPSS**, razón por la cual acuden a él para incorporar en sus consolas de programación las librerías específicas que aquellas demandan, y que numerosos miembros de la comunidad científica que participan en el desarrollo del entorno **R** han puesto libremente al alcance de todos los humanos, al renunciar a la propiedad intelectual de sus productos de programación.

Es importante anticipar que, aunque se está hablando de recursos estadísticos, la utilización de **RStudio** no se limita a las *investigaciones aplicadas*, ni tampoco a las que siguen únicamente el *enfoque cuantitativo*. Junto a la condición de *software libre* y su amplia capacidad de actualizarse con las técnicas estadísticas más avanzadas, este entorno informático tiene la extraordinaria virtud de servir también de herramienta al procesamiento de códigos y constructos de las llamadas *investigaciones teóricas* (incluyendo las de corte histórico), así como de datos provenientes de la aplicación de instrumentos semi-estructurados o no estructurados, propios de las investigaciones que siguen el *enfoque cualitativo*, al capturar, organizar y procesar datos obtenidos directamente de la práctica social.

Lo dicho hasta aquí puede asociarse tan solo a una cuestión técnica, a la utilidad de cierta herramienta para una determinada acción científica. Y en parte lo es, pues es un hecho conocido que la ciencia sin una técnica apropiada no puede

desarrollarse; basta recordar que la Astronomía no fue la misma antes que después de Galileo Galilei y de su mejoramiento del telescopio, como mismo que la medicina moderna no pudiera efectuar las asombrosas contribuciones que hoy realiza a la salud humana sin el sistema instrumental que la asiste.

Sin embargo, aquí se está queriendo hablar de mucho más. La intención del presente trabajo es presentar a **RStudio** como una oportunidad insuperable para los investigadores educativos cubanos de disponer de un potente y eficaz software que no le costará a sus instituciones ni siquiera un céntimo por concepto de licencia y que, al mismo tiempo, tiene la enorme virtud de contener en sí mismo recursos de programación necesarios para cubrir las necesidades de puntos clave de las investigaciones sociales y humanísticas, a saber: la determinación del diseño teórico-metodológico, la conformación de un marco teórico-referencial apropiado, la elaboración de instrumentos de investigación válidos y confiables, y el análisis lógico y estadístico de los datos derivados de su aplicación, pero también la confrontación colegiada de las conclusiones derivadas de aquellos, la difusión de los resultados científicos obtenidos y su introducción eficaz en la práctica educativa.

Si bien la primera es evidente, axiomática, esta última idea está por sustentar. Luego, una prioridad del presente trabajo será, primero, argumentar hasta qué punto esos *nodos esenciales* del proceder investigativo (en el campo educacional cubano) requieren ser mejorados y, segundo, fundamentar en qué medida **RStudio** (y por extensión **R**) constituye una herramienta instrumental eficaz para perfeccionarla.

Desarrollo

En la **Tesis** en opción al grado científico de Doctor en Ciencias (Torres, 2016), titulada “Retos de la investigación educativa cubana actual. Aportes a su tratamiento”, este autor aisló lo que denominó allí *retos* a enfrentar por la comunidad cubana de investigadores en el campo educacional, con vistas a poder lograr el reclamado mejoramiento de sus procederes metodológicos (Rubio, 2000); (Torricella, Van Hooydonk & Araujo, 2000); (Bermúdez & Rodríguez, 2008); (Arencibia-Jorge & de Moya, 2008); (Chirino, 2009); (Valledor et al., 2009); (Cruz,

2009); (Pérez et al., 2009); (Ortiz, González-Guitián, Infante & Viamontes, 2010); (Arnaiz & García-Rodríguez, 2011); (Mainegra & Miranda, 2012); (Cruz, Escalona & Téllez, 2014); (García-Cespedes, Montejo & Carvajal, 2014); (Ortiz, 2015); y, en consecuencia, la elevación de la calidad de sus resultados científicos.

Esos **retos** identificados fueron los siguientes:

- la correcta utilización de los enfoques de investigación *cuantitativo*, *cualitativo* y *mixto*,
- el empleo combinado de la *operacionalización* de las variables y el *camino ascendente* desde los datos,
- el tratamiento de la *validación práctica* de los resultados de las investigaciones aplicadas,
- la *visibilidad internacional* de los investigadores educativos cubanos,
- el fortalecimiento de la *relación entre comunidades cubanas* de investigación educativa,
- el tratamiento del concepto de *impacto social* de la investigación educativa cubana.

Se trata de una situación preocupante, que proyecta una inquietante *sombra* sobre el pilar fundamental de la investigación científica: la **objetividad** de sus resultados. En efecto, el imprescindible rigor del proceder científico se ha visto afectado tanto desde la *perspectiva epistemológica* (o sea desde la relación clave *investigador(es)-objeto de estudio*), como también desde la no menos importante *arista sociológica* (entiéndase, desde la relación *investigador-investigadores-sociedad*). Sin los altos niveles de credibilidad que solo pueden posibilitar el rigor metodológico y el cuestionamiento razonable entre pares de los procedimientos, técnicas y productos de la actividad científica, la investigación educativa no puede aspirar al cumplimiento de sus elevados compromisos profesionales y políticos contraídos con la sociedad cubana, ni aspirar a que sus resultados sean tenidos en cuenta y generalizados por sus instituciones en la práctica educativa y social.

Complementa el sustento principal la Tesis, otras observaciones críticas subordinadas a aquel, y que igual es imprescindible que sean tenidas en cuenta, en

tanto relación de lo *táctico* con lo *estratégico*. Una parte importante de estos otros aspectos corresponden a falencias de carácter metodológico, pero también a limitaciones en el *manejo de recursos lógicos, estadísticos, informáticos y comunicacionales*, requeridos para alcanzar una actividad investigativa profunda y eficaz.

Así, la Tesis evidenció dificultades con la **utilización de fuentes bibliográficas** afines a un mismo punto discursivo, al igual que con el consecuente y abarcador **análisis crítico de las posiciones teóricas** en ellas contenidas (al punto de evidenciarse una tendencia al incremento de las obras consultadas, pero no así en el número de referencias bibliográficas y de citas extraídas de ellas, tan necesarias estas últimas para poder polemizar en torno a los diferentes componentes del *marco teórico-referencial*, por ejemplo).

Igualmente, se apreció la existencia de una cantidad no despreciable de insuficiencias relacionadas con la **construcción de los instrumentos de investigación** (acorde con el *enfoque investigativo* elegido, pero también producto de un manejo deficiente de la *operacionalización* de las *variables principales* de la investigación, en particular con el paso directo de un *indicador* a su respectivo *reactivo*). De forma remarcada, se soslaya la cuestión del **escalamiento de los reactivos** de los instrumentos en la investigación educacional cubana actual.

Especialmente afectado se apreció, entre los más de mil trescientos reportes de investigación revisados, la cuestión del **manejo estadístico** del proceder investigativo. No solo por un marcado desconocimiento de los diferentes tipos de Estadística (*Descriptiva*, *Inferencial* y de *Análisis Multivariado*) y de sus posibilidades de aplicación, de acuerdo con los tipos de variables de investigación, sino también como resultado de la existencia de no pocos y de serios errores conceptuales en torno a la conveniencia o no de la utilización de *muestras estadísticas* y a la deficiente **selección de muestras estadísticas** o al predominio de las de carácter intencional, con la pretensión igual de continuar haciendo inferencias estadísticas de los *estadígrafos muestrales* a los *parámetros poblacionales*, como si se trataran de *muestras aleatorias y representativas*.

En lo que se refiere a los **análisis estadísticos de los datos** extraídos de la práctica escolar, las insuficiencias y los errores manifiestos no son menos. Aun cuando desde el discurso científico se reconoce cada vez más que los fenómenos educativos poseen un marcado *carácter multifactorial*, los procesamientos siguen siendo muy elementales (casi siempre acotados al *cálculo porcentual*), y las inferencias lógicas (a modo de conclusiones investigativas) están poco sustentadas en la interpretación de los datos.

Otra singularidad apreciada durante el estudio diagnóstico practicado, con los centenares de tesis de maestría y de doctorado, así como de reportes de resultados científicos de proyectos de investigación, radica en el pobre desarrollo de **habilidades informáticas** relacionadas con el manejo de gráficos ilustrativos de las relaciones estadísticas y/o bibliométricas que pudieran aflorar de los estudios investigativos practicados. Así como los análisis numéricos se suelen reducir al cálculo de por cientos, los gráficos estadísticos o sus similares, para las relaciones de categorías conceptuales y citas bibliográficas, quedan circunscritos al empleo de simples representaciones de *barras* y *pasteles*, como norma. Ello repercute, al mismo tiempo, en la **calidad estética y comunicacional de los reportes de investigación**, reduciendo de forma considerable la comprensión de los *hallazgos científicos* expuestos y de su potencial impacto sobre el *objeto de estudio* o sobre los *sujetos de la investigación*, según el caso.

Como podrá apreciarse, el problema de la calidad de la *investigación educativa cubana actual* es mucho más complejo que lo que puede inferirse, a primera vista, de la presentación de los seis *retos* generales recogidos en la Tesis; desafíos que son ya, por sí solos, enormes. De manera que este complejo panorama sugiere una selección razonada de las *herramientas investigativas* más pertinentes y racionales a generalizar. Pues bien, la *buena noticia* es que **RStudio** es un magnífico candidato para ello.

Imposible de ilustrar la aseveración anterior en un único artículo científico, se intentará a continuación iniciar, al menos, un acercamiento a esa exigencia, a través

de la ilustración de diversas situaciones investigativas reales (previamente desarrolladas por el autor) en las que el **entorno informático R** (a través de su herramienta auxiliar **RStudio**) muestra sus potencialidades para intervenir positivamente en esa impostergable *cruzada* global.

Para comenzar, se presentará un ejemplo de análisis semántico de palabras clave en la conformación de un *marco teórico-referencial*, tras la elaboración del *diseño teórico-metodológico* de la investigación. La representación gráfica obtenida estará acompañada del **scrip** (o secuencia de programación en **RStudio**) que dio lugar a la misma, a manera de una primera familiarización con el entorno **R**. El *bloque de códigos de R* o **chunk**, en cuestión, y el producto de su *corrida* son, respectivamente, los siguientes:

```
library(tm)
tesis=VCorpus(DirSource("C:/WordCloud", encoding="UTF-8"), readerControl=list(language="spa"))
inspect(tesis)
tesis=tm_map(tesis, tolower)
tesis=tm_map(tesis, removePunctuation)
tesis=tm_map(tesis, removeWords, stopwords("spanish"))
tesis=tm_map(tesis, removeWords, c("marxismo","epistemología"))
library(wordcloud)
wordcloud(tesis[[1]], scale=c(2.5,0.5), max.words=100, rot.per=0.25, colors=brewer.pal(8, "Dark2"))
```

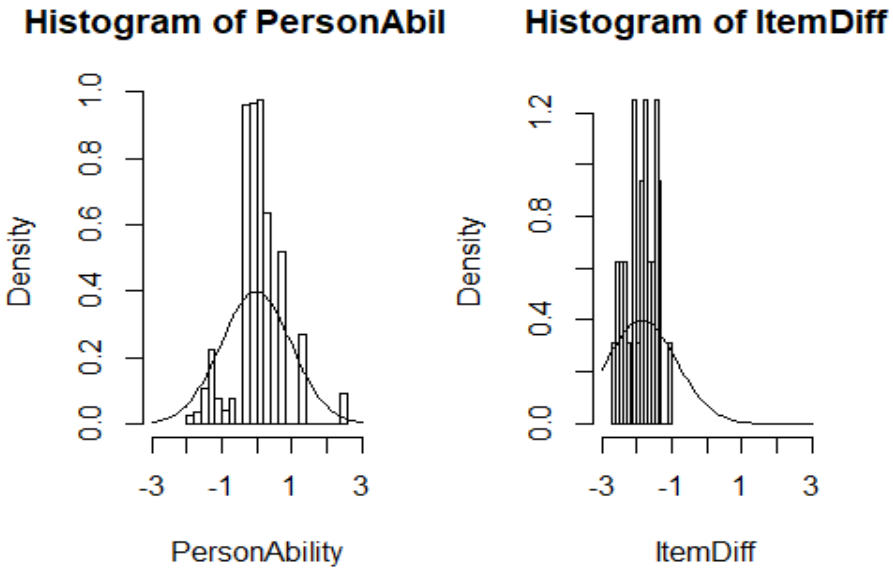
Este resultado se ha obtenido a través de la programación en **RStudio** de un algoritmo para conformar una **nube de palabras** representativa del contenido de la primera parte de la ya referida Tesis de segundo grado del autor del presente trabajo. Obsérvese que con este gráfico se pudieran hacer valoraciones preliminares acerca de las principales *categorías* del marco teórico-referencial asumido (a manera de empleo de la modalidad investigativa de *teoría fundamentada*, del *enfoque cualitativo*). Al mismo tiempo, se puede prestar la atención a los autores más referidos en esa parte del documento, lo cual anticipa la posibilidad de trabajar también *análisis bibliométricos*, otra arista importante de la actividad científica.

En relación con la construcción de los instrumentos de investigación, se ilustrará el *escalamiento de los puntajes* de una *variable-producto* medida por él, con la transformación de la escala asociada a la *Teoría Clásica del Test* (TCT) por otra representativa de la *Teoría de Respuesta al Ítem* (TRI). Se trata de una cuestión trascendente para la *confiabilidad* de un instrumento de investigación, pues en la

escala clásica (generalmente acotada entre cero y cien) no es posible separar el nivel de desarrollo de la *habilidad latente* en el educando, del grado de *dificultad del reactivo* del instrumento. Así, en el caso de *pruebas cognitivas*, por ejemplo, al *pilotearlo* el instrumento resultará *fácil* si la muestra a la que se le administra está compuesta predominantemente por estudiantes de alto rendimiento, pero ese mismo instrumento será considerado *difícil* si lo que prevalece en el colectivo seleccionado son estudiantes de bajo rendimiento.

Se ilustrará a continuación la conversión de la escala de evaluación de los puntajes y cómo **RStudio** (a través de la *librería* TAM) logra separar objetivamente la *habilidad latente* de la *dificultad del reactivo*. El *bloque de códigos de R* o *chunk* y el resultado de su *corrida* son ahora los siguientes:

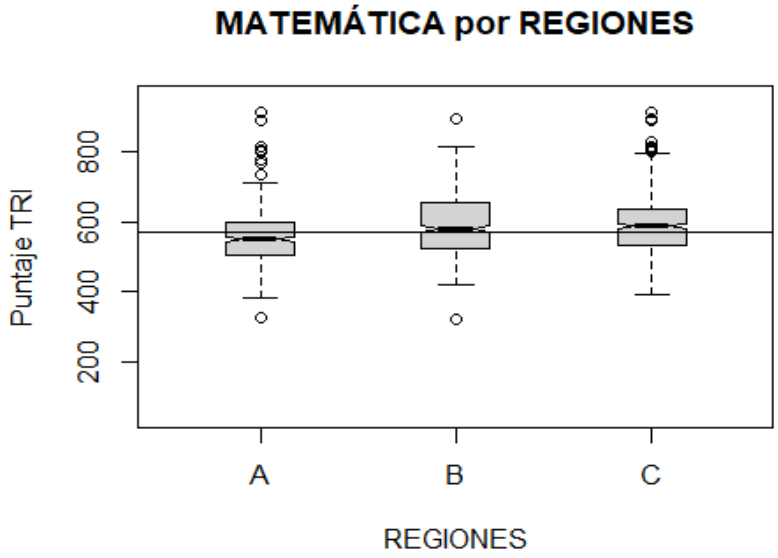
```
setwd("C:/datos")
dat.csv <- read.csv("prueba.csv")
attach(dat.csv)
na.omit(dat.csv)
library("TAM")
mod1 <- tam(dat.csv)
ItemDiff <- mod1$xsi$xsi
Abil <- tam.wle(mod1)
Abil
PersonAbility <- Abil$theta
par(mfrow=c(1,2))
hist(PersonAbility,xlim=c(-3,3),breaks=20, prob=TRUE)
curve(dnorm(x, mean=mean(PersonAbility)), add=TRUE)
hist(ItemDiff,xlim=c(-3,3),breaks=20, prob=TRUE)
curve(dnorm(x, mean=mean(ItemDiff)), add=TRUE)
```



Como puede apreciarse, se ha logrado separar el comportamiento de la *habilidad latente* de los educandos, en relación con el constructo que está siendo objeto de evaluación (*PersonAbility*), de los *niveles de dificultad* de los componentes del instrumento (*ItemDiff*); se trata de una colección de reactivos de bajo nivel de dificultad pues, a diferencia de la habilidad, sus puntajes en la nueva escala (correspondiente a la TRI) se distribuyen muy a la izquierda de cero (ente -2.7 y -1). Un análisis docimológico de cada uno de esos ítems, por separado, podría proporcionar más información en cuanto a la conveniencia o no de modificar el instrumento, proceso que es posible realizar también con la *librería* TAM. Es oportuno aclarar que estos análisis no son válidos solo para instrumentos que constituyen *pruebas de logro cognitivo*, como a veces se cree, sino también para otros, como los *cuestionarios*.

Hasta aquí, se ha podido ilustrar las potencialidades de **RStudio** para apoyar etapas iniciales de la actividad científica, como el avance hacia la conformación del *marco teórico-referencial* y la construcción y calibración de los *instrumentos de investigación*. A continuación se ejemplificarán sus fortalezas en torno a la realización de **análisis estadísticos** de los datos obtenidos con los instrumentos, tanto los simples como los de mayor alcance.

Se comenzará con la utilización de una gráfica *de caja y bigotes* (o **boxplot**) para mostrar el comportamiento de los resultados de una cierta prueba de logro (en este caso de Matemática) por regiones geográficas. El resultado gráfico es el siguiente:



Aparecen en el gráfico las *medianas* de las submuestras que conforman cada región (representadas por las líneas oscuras situadas en el interior de cada *caja*), así como las *puntuaciones máximas* y *mínimas* alcanzadas en cada una de ellas (extremos de los *bigotes*), excepto los puntajes que constituyen *outliers* (o *valores atípicos*), pues se separan de la mediana a 1.5 veces la altura de la *caja intercuatílica*; estos casos atípicos aparecen representados con pequeños círculos situados por encima y/o por debajo de los *bigotes*.

Así, desde la **Estadística Descriptiva** es posible señalar que las regiones aquí denominadas “B” y “C” alcanzaron *medianas* ligeramente superiores a la *mediana global* de la muestra (simbolizada con la línea negra paralela al eje “x”), pero con mayor dispersión en los puntajes de su submuestra (pues sus *bigotes* son *más largos* que los de la región “A”). En cambio, si se desea explorar si las diferencias entre las *medianas* de las regiones son *estadísticamente significativas* (ahora, desde la **Estadística Inferencial**), no es necesario realizar la *prueba de hipótesis* correspondiente, pues bastó agregar el *parámetro* **notch** a la función *boxplot* en el

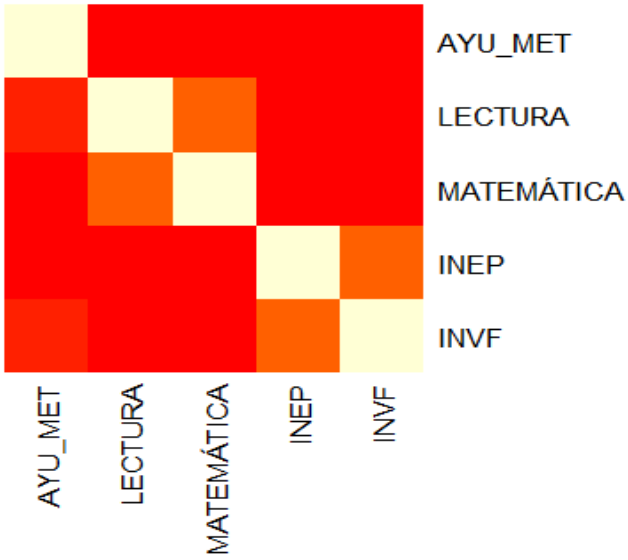
chunks anterior.

Como resultado de ello, a la altura de la *mediana* de cada *caja* aparece una muesca; puesto que las muescas de las regiones "A" y "B" (como mismo las de las regiones "A" y "C") no se solapan, puede afirmarse, con una *alta probabilidad*, que dichos pares de *medianas* son significativamente diferentes. Por el contrario, la *diferencia significativa* no está asegurada entre las *medianas* de las regiones "B" y "C".

Los ejemplos hasta aquí tratados refuerzan la idea de que el entorno informático **R** destaca por la plasticidad y potencia de sus representaciones gráficas. Otros tipos de gráficos potentes, generados fácilmente desde **R**, son el **heatmap** (o *mapa de calor*) y el **dendograma**, ambos de utilidad para representar relaciones cercanas entre *variables* o entre *unidades de análisis* de una muestra.

Así, por ejemplo, si se dispone de una *base de datos* (o **data frame**) conformada por los resultados de la aplicación de *pruebas de logros del aprendizaje* y de *cuestionarios* sobre *factores asociados* a dichos logros (de manera que estos últimos constituyan un acercamiento a la explicación de las marcadas variaciones que se suelen presentar entre los puntajes de los *logros del aprendizaje* de los educandos), entonces un **heatmap** contentivo de algunas de las variables representativas de cada uno de esos dos grupos de instrumentos puede ayudar a esclarecer los niveles de **correlación bivariada** que existe entre ellas, sobre la base de hipótesis asumidas desde la teoría pedagógica de partida.

En la representación que sigue se ha indagado sobre los niveles de asociación existentes entre los puntajes de los estudiantes de la muestra en las *pruebas de logro* de Lectura y Matemática, y los valores añadidos de: un *índice del nivel educativo de sus padres* (*INEP*), otro *del nivel de vida de su familia* (*INVF*), y uno tercero del *nivel de la calidad de la preparación metodológica* de los docentes que le imparten clases (*AYU_MET*), según el criterio del director de la escuela.

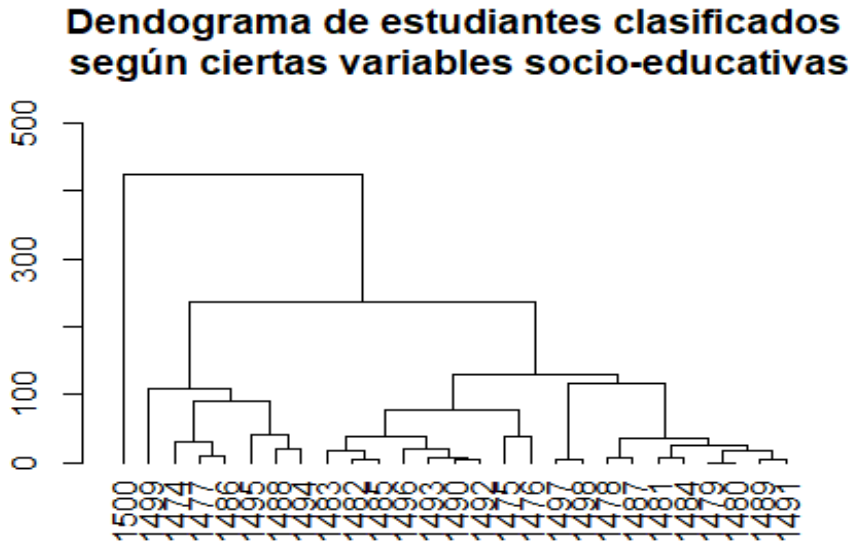


Nótese que, si bien se aprecia cierto nivel de *correlación estadística* entre los puntajes de Lectura y los de Matemática, por un lado, y entre los *índices del nivel de vida familiar* y *del nivel educativo de los padres*, por el otro (pues la intersección de sus respectivas filas y columnas reportan cuadrados de colores claros), no sucede lo mismo en las intersecciones de los puntajes de Lectura con ambos índices familiares, como tampoco en las de los puntajes de Matemática con ellos. Es decir, el **heatmap** está indicando que los resultados de las *pruebas de logros del aprendizaje* aplicadas están poco asociados al llamado **capital familiar** y, por tanto, la elevación de sus resultados parece depender esencialmente de la intervención pedagógica que se realice, y no de las condiciones socio-familiares, externas al proceso docente-educativo.

También puede observarse que la *calidad de la preparación metodológica* que se ha venido realizando con los docentes de esos estudiantes tampoco está fuertemente asociada a los *logros del aprendizaje* en esas asignaturas, fortaleciendo la necesidad de un mejoramiento del funcionamiento de las instituciones escolares de la muestra de interés.

En el caso del ejemplo de **dendograma**, se hará foco en la cercanía (o distanciamiento) entre las *unidades de análisis* (por ejemplo, entre estudiantes o entre centros educativos). Este otro tipo de representación gráfica se apoyará, en

consecuencia, en el cálculo de la *matriz de distancias euclídeas*. Al extraer una **submuestra** de la *base de datos* anterior (para reducir el número de *unidades de análisis* a considerar, y con ello de la complejidad del gráfico), se obtiene la siguiente representación.



Obsérvese en ella que los estudiantes de la submuestra seleccionada han sido agrupados ordenadamente en **clústers jerárquicos**, en tanto sus medidas de *distancia euclídea* resulten más parecidas. Así, los estudiantes de las filas números 1479 y 1480 son, sin discusión, los de comportamiento más próximo (en relación con las variables consideradas en el nuevo *data frame*), como mismo el educando de número de orden 1500 es el que más se diferencia de los restantes, en ese colectivo.

Si bien los *gráficos estadísticos* generados por **RStudio** son muy importantes, los lectores no deberían formarse el criterio de que los resultados analíticos en **R** lo son menos. Este entorno informático, como se ha dicho, es un *paquete estadístico* sumamente versátil y actualizado, luego pueden desplegarse con él novedosos y potentes *análisis estadísticos*, que resultan ser, además, especialmente útiles para las investigaciones sociales y humanísticas, dado que el número de variables a

considerar en ellas son muchas y, mayoritariamente, de naturaleza *cualitativa*.

Para ilustrar la aseveración anterior se incorporarán al presente trabajo dos ejemplos de tipo analítico: uno referido a un **Análisis de Varianza** y otro a los novedosos **Modelos Jerárquicos Lineales**.

Se desplegará muy brevemente un *análisis de varianza de un factor* (**ANOVA**), en el que el objetivo será verificar si las *medias* de tres *categorías* de un **factor** difieren significativamente entre sí (con lo cual se estaría afirmando que el *factor* es independiente de la *variable dependiente* de la función que modela matemáticamente la situación que está siendo objeto de estudio). Concretamente, se indagará si las puntuaciones obtenidas por los estudiantes en la *prueba de logro* de Matemática, en la muestra recogida en el *data frame* "prueba2", son independientes de las *regiones* del país a la que ellos corresponden (recuérdese del ejemplo del *boxplot*: "A", "B" y "C"). Para ello basta agregar dos líneas de códigos más al *chunk* anterior.

```
REGIONvsMAT <- lm(MATEMÁTICA~REGIONES, data=prueba2)
anova(REGIONvsMAT)
## Analysis of Variance Table
##
## Response: MATEMÁTICA
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## REGIONES     2  592647  296323  44.965 < 2.2e-16 ***
## Residuals 1497 9865396    6590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La tabla del *análisis de varianza* realizado está señalando que el *p-valor* asociado a la *distribución F* (2.2e-16) es sumamente menor que el *nivel de significación estadística* asumido (0.001), por lo que puede afirmarse que existen diferencias de consideración entre las *medias* de las puntuaciones de Matemática, en relación con las tres regiones de la población de la que se extrajo la muestra. Claro, se está asumiendo que se cumplen los supuestos de un **ANOVA**: la *independencia de las*

mediciones realizadas, la *normalidad de la distribución de los residuales* y la *homogeneidad de las varianzas* de las categorías del *factor*, aspectos estos últimos para los que el **entorno informático R** también proporciona valiosos recursos estadísticos, tanto analíticos como gráficos.

El **ANOVA**, asociado en el ejemplo anterior a un estudio de corte *ex-post-facto* (como los **ERCE** del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación, de la **UNESCO**), es particularmente útil también para los **estudios pedagógicos experimentales**, donde el *factor* está compuesto por **niveles de tratamiento** de la *variable independiente*, en vez de *categorías*.

El otro ejemplo de tipo analítico que se pretende presentar, como se dijo, se refiere a la utilización de los **Modelos Jerárquicos Lineales**; se trata de un grupo de recursos relativos al **Análisis Multivariado** de cierto nivel de complejidad técnica, pero que son especialmente útiles para la comprensión a fondo de fenómenos sociales complejos, donde intervienen simultáneamente numerosas **variables predictivas** de una o de varias **variables de producto** y, al mismo tiempo, las unidades de análisis asumen una **estructura anidada** (Bliese, 2016); un caso típico: ¡los resultados de los *logros en el aprendizaje* (o del desarrollo de determinada *habilidad socioemocional*) en los escolares de cierto grupo de instituciones educativas!

¿Por qué? Pues, porque sobre esos *logros* de los educandos, que es una *variable de producto* (del proceso docente-educativo), han estado interviniendo decenas de *variables predictivas* de índole familiar, social, aúlico y escolar. Pero, además, porque esos educandos ven modificadas sus características personales previas, en alguna medida, como resultado de su agrupamiento en colectivos escolares (aulas), que después se integran a colectivos mayores (escuelas); es decir, porque conforman una **estructura jerárquica**.

Así los **Modelos Jerárquicos Lineales** permiten una mayor aproximación al estudio de los resultados del proceso educativo que otras herramientas estadísticas, a la vez que posibilita discriminar las *variables predictivas* de mayor capacidad explicativa y también, para más potencia aún, la *fuerza* con que ellas inciden sobre

la *variable de producto* de contraste, o sea se convierte en una **evaluación de impacto** precisa. Por tanto, un recurso investigativo tan aportador merece el fuerzo técnico que demanda.

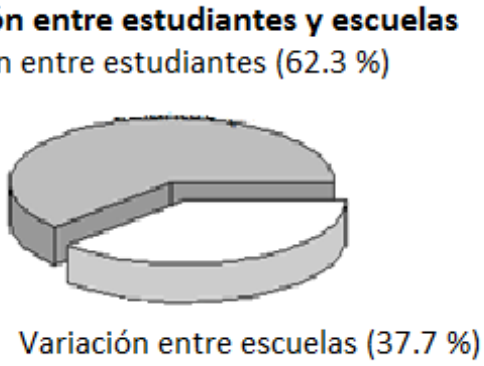
Lo primero que demanda un *análisis jerárquico lineal* es la verificación de que la *estructura anidada* de las unidades de análisis *particiona* la *varianza global* de los datos de la *variable de producto*; es decir, que esta última no está determinada solo por la *varianza* de los atributos individuales (o *primer nivel*), sino que también la agrupación de las unidades de análisis (o *segundo nivel*) genera una *varianza* asociada a ella distinto de cero. Este primer modelo a construir se le denominado **modelo nulo**, dado que no considera todavía ninguna *variable predictora* de la *variable de producto*.

Para ilustrarlo se seguirá empleando el mismo *data frame* (o base de datos) de los últimos ejemplos desarrollados; el *primer nivel* (de atributos individuales) estará representado por la variable "INEP" (*índice del nivel educativo de los padres de los educandos*), la variable de producto será "MATEMÁTICA" (o sea los puntajes de los estudiantes de la muestra en la *prueba de logros del aprendizaje* de esa asignatura) y las unidades de agrupamiento (en un *segundo nivel*) serán las escuelas (representadas por la variable "IdESC"). También se han generado, para reducir las desviaciones en el análisis, las variables adicionales "meanINEP" y "centINEP", referidas al *promedio del índice del nivel de vida de los padres por centro educativo* y, sobre su base, el *valor centrado* del *índice del nivel de vida de los padres* de cada estudiante, respectivamente.

El *bloque de códigos de R* (o *chunk*) con el que se generará este primer resultado parcial (de *modelo nulo*) cierra con el cálculo del *coeficiente de correlación intraclase* (por sus siglas en Inglés: **ICC**), que representa la proporción de la *varianza total* de la *variable de producto* que corresponde al *nivel 2* (o sea, no al individual, sino al de anidamiento). Adicionalmente, nótese que lo que se pretende hacer aquí es precisamente un **ANOVA factorial**, donde los componentes del *factor* son los códigos de identificación de las escuelas del *data frame*.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: MATEMÁTICA ~ 1 + (1 | IdESC)
##    Data: prueba3
##
## REML criterion at convergence: 17080.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0982 -0.6235 -0.1230  0.5087  4.3111
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  IdESC    (Intercept) 2594     50.93
##  Residual             4282     65.43
## Number of obs: 1500, groups:  IdESC, 178
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  569.325      4.402   129.3
## [1] 0.3772542
```

Como habrá podido apreciarse en la devolución estadística anterior, el resultado del cálculo del **ICC**, para los resultados de los *logros del aprendizaje en Matemática* de los 1500 estudiantes, de las 178 escuelas de la muestra seleccionada, representa un valor mayor que cero (0.3772542); puede afirmarse, entonces, que el 37.7% de la *varianza total* de esos puntajes es atribuible a variaciones entre escuelas y que, por tanto, no toda la *varianza total* depende de las diferencias personológicas de los estudiantes. En términos gráficos sería como sigue.



Ello significa un estímulo para continuar perfeccionando este *modelo lineal* inicial, representado por la **ecuación de efectos mixtos** siguiente:

$$\mathbf{Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}}$$

en el que $\mathbf{Y_{ij}}$ representa las variaciones de los puntajes en Matemática de cada estudiante (**i**) dentro de cada escuela (**j**), $\boldsymbol{\gamma_{00}}$ la *gran media*, $\mathbf{u_{0j}}$ es la desviación de los promedios de las escuelas de la *gran media* y $\mathbf{r_{ij}}$ la desviación de los puntajes de cada estudiante del promedio de su escuela.

Tal perfeccionamiento es posible a través de la incorporación gradual de *variables predictoras*, tanto del *primer nivel* (por ejemplo "meanINEP"), como del *segundo* (dígase, "AYU_MET"). Por un problema de extensión de este trabajo, evidentemente de carácter introductorio a la temática, no se desplegará el resumen estadístico de la *corrida* de esa segunda *ecuación de efectos mixtos*; en su lugar se expondrá solo el *bloque de códigos de R* correspondiente.

```
results2M=lmer(MATEMÁTICA~1+ meanINEP+ AYU_MET + (1|IdESC), data = pr
ueba3)
##summary(results2M)
```

Un tercer modelo lineal, especialmente explicativo, se logra cuando se utiliza el recurso de *interacción entre niveles* (**cross-level interaction**) que actúa como especie de un **ajuste de variable** (o sea, de la relación entre dos variables a través de una tercera variable). En el ejemplo que se viene siguiendo se interaccionará la *variable predictiva* consistente en el *nivel de ayuda metodológica recibida por los docentes de los estudiantes* ("AYU_MET") con la variable del *primer nivel* referida

al *índice del nivel de vida de los padres* de cada estudiante, centrado en el promedio de la escuela ("cenINEP").

El efecto que se pretende investigar aquí es: *cuánto explica el nivel de ayuda metodológica recibida por los docentes, el resultado en los logros en Matemática de sus estudiantes, de acuerdo con el nivel educativo de los padres*; se trata, prácticamente, de un análisis personalizado, o como se suele decir en el argot pedagógico: **"atendiendo a las diferencias individuales"**. Por su importancia, en este caso si se desplegará completamente el *chunk* y los resultados del *análisis jerárquico lineal*.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## MATEMÁTICA ~ meanINEP + AYU_MET + centINEP + AYU_MET * centINEP
+
##     (1 + centINEP | IdESC)
##    Data: prueba3
##
## REML criterion at convergence: 17056.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0947 -0.5977 -0.1333  0.5053  4.1433
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  IdESC    (Intercept) 2640.78  51.388
##           centINEP      59.41   7.708   0.17
##  Residual             4178.18  64.639
## Number of obs: 1500, groups:  IdESC, 178
##
## Fixed effects:
```

```
##                 Estimate Std. Error t value
## (Intercept)     561.64156   21.58738  26.017
## meanINEP          3.56167    3.91258   0.910
## AYU_MET          -1.49319    4.71795  -0.316
## centINEP          3.19622    6.36786   0.502
## AYU_MET:centINEP  -0.08394    1.80273  -0.047
##
## Correlation of Fixed Effects:
##             (Intr) mnINEP AYU_MET cnINEP
## meanINEP    -0.633
## AYU_MET     -0.737 -0.015
## centINEP     0.047 -0.003 -0.056
## AYU_MET:INE -0.046  0.004  0.058  -0.968
```

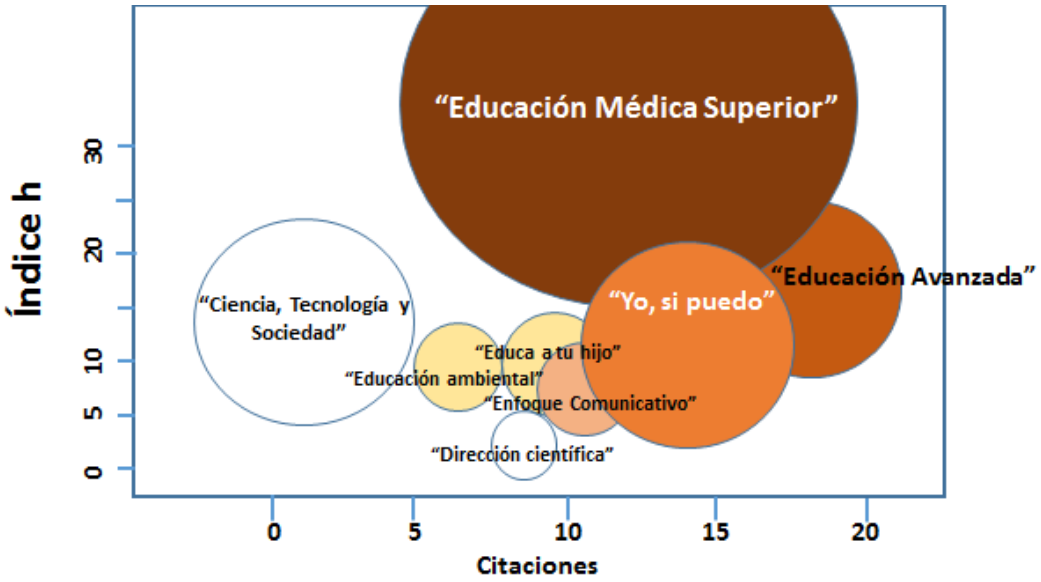
Además de ratificar la *ecuación de efectos mixtos* de este tercer modelo lineal y el criterio empleado por la función utilizada para procesarlo (REML), aparece un resumen estadístico de los *residuos*, de los *efectos aleatorios* y de los *efectos fijos obtenidos*, además de la matriz de correlación entre estos últimos. El interés ahora, aquí, se centrará en las magnitudes de los efectos fijos de la *variable predictora* "AYU_MET" sobre los resultados en "MATEMÁTICA", tanto de forma aislada como interactuando con el índice del nivel educativo de los padres de los estudiantes, centrado en el promedio de la escuela ("AYU_MET:centINEP").

Nótese que ambos valores están muy próximos a cero (-1.49 y -0.08), e incluso son negativos. El modelo lineal está diciendo así que, para esa muestra de estudiantes y escuelas, *la ayuda metodológica que reciben los docentes de las escuelas de sus directores*, como norma, *no tiene un efecto positivo sobre los resultados del aprendizaje en Matemática de sus estudiantes*, y más aún: que *esa tendencia no varía con independencia de que los estudiantes provengan de núcleos familiares con un ambiente educativo y cultural favorable o no*. En pocas palabras, que la intervención científico-metodológica de las estructuras de dirección inmediatas a las escuelas no debería posponerse, si se desea modificar para bien la situación.

Claro, el proceso educativo es marcadamente multifactorial; por tanto, las otras *variables potencialmente predictoras* del *data frame* debieran también ser analizadas. Por suerte, o mejor gracias al **entorno informático R**, no hay que partir de cero, pues los **scripts** (o guiones de programación) ya están elaborados y solo se trata de sustituir unas variables por otras, incluyendo la repetición de las *corridas* con otras asignaturas evaluadas.

Así, se ha podido ilustrar el empleo del **entorno informático R** en momentos de la fase de *elaboración de un marco teórico-referencial*, de la fase de *elaboración de los instrumentos de investigación* y, con varios ejemplos, vistos tanto desde la perspectiva gráfica como desde la analítica, de la fase *del análisis estadístico de los datos*. Sin embargo, en (Torres, 2016) se ha insistido en que el proceso investigativo no se limita a la relación principal *investigador-objeto de estudio*, sino que incluye también las acciones asociadas a la, no menos importante, relación *investigador-investigadores-sociedad*.

En consecuencia, se incluirá otro ejemplo referido a las potencialidades de **R** en torno al desarrollo de actividades correspondientes a esa otra arista del proceso investigativo. Concretamente, se ilustrará el empleo de **gráficos de burbujas** para el análisis de la **productividad bibliométrica** de comunidades científicas nacionales, representativas de diferentes líneas de investigación en el campo educacional. En el mismo, los tonos de colores más oscuros representarán una cantidad mayor de *documentos producidos* por las diferentes líneas de investigación identificadas, los desplazamientos más a la derecha del eje horizontal significarán mayor número de *citas recibidas*, y más hacia arriba en el eje vertical un mayor *índice h*, de productividad; finalmente, mayor longitud del diámetro de los círculos consignará un promedio mayor de *citas anuales*. Véase a continuación los resultados de la ejecución del *chunk*.



La representación del tipo **bubble** (o *gráficos de burbujas*) obtenida está indicando que la línea "Educación Médica Superior" es, por mucho, la de mejores resultados pues, con una cantidad considerable de documentos y de citas anuales, ha conseguido una acumulación mayor de citaciones y de productividad bibliométrica, en términos de "Índice h". Le continúan, en ese orden, las líneas: "Educación Avanzada", "Yo, sí puedo" y "Educa a tu hijo". No menos meritoria ha sido la producción documental de "Ciencia, Tecnología y Sociedad", que ha alcanzado un "Índice h" a pesar de no tener muchas citas, ni documentos producidos.

Antes de concluir, es importante destacar que el **entorno informático R** es útil hasta para la redacción misma de los reportes de las investigaciones realizadas. Ello es posible a través de la composición de ficheros **R Markdown**, que permiten la inclusión, en la redacción del texto (bien en Word, en PDF o en HTML), de *fórmulas matemáticas*, de los *bloques de códigos R* (o *chunks*) y de los *chunks en modo línea*.

De hecho, el presente trabajo ha sido redactado en un fichero del tipo **R Markdown**. En algunas de sus partes ya se han mostrado los *bloques de códigos R* que se han empleado para generar ciertos análisis estadísticos y representaciones gráficas de diferentes tipos, pero debe revelarse que con su ayuda también se han introducido fórmulas matemáticas; ese es el caso de la que modeló la ecuación lineal de efectos

mixtos, arriba representada, la que demandó de una compleja sintaxis no visible en este documento, tras la *corrida* del fichero.

La ventaja principal de componer ficheros *R Markdown* al redactar los informes de investigación, en vez de escribirlo desde un procesador de textos e incluirle las ecuaciones, gráficos y análisis estadísticos generados en **R**, es que permite una mayor facilidad en su re-elaboración, en caso de tener que hacer cambios en los datos primarios. En todo caso, un fichero *R Markdown* se convierte, automáticamente, en un valioso **script** (o secuencia de comandos de un programa de computación) para futuros trabajos científicos similares.

Conclusiones

Con el presente trabajo se ha cumplido el objetivo de realizar una aproximación al **entorno informático R**, visto este no solo como una poderosa herramienta auxiliar de la investigación científica contemporánea, sino también como una valiosa oportunidad de asistir al reclamado perfeccionamiento de la **investigación educativa cubana**, en la actualidad.

Con el despliegue de algunos ejemplos de procesamientos estadísticos en **RStudio** se ha podido ilustrar cómo ese conjunto integrado de **librerías** e instrucciones informáticas puede resultar de utilidad en diferentes fases del proceso investigativo (como en la conformación de un *marco teórico-referencial*, en la elaboración de *instrumentos de investigación*, en la construcción de *gráficos* y el desarrollo de *análisis estadísticos*, en la realización de *estudios bibliométricos* y durante la redacción de los *reportes de investigación*, entre otros). Se trata, justamente, de acciones en torno a las cuales se han identificado falencias en una parte no despreciable de investigadores cubanos, del sector educativo.

Al mismo tiempo, hay que tener en cuenta que **R** se sustenta en la filosofía del *software libre*, algo extraordinariamente transcendente para Cuba, un país de limitados recursos financieros y sometido a un permanente bloqueo económico, comercial y financiero por parte de la principal potencia mundial, desde hace casi 60 años.

Ojalá que la *comunidad cubana de investigadores educativos* se percate de todas estas facilidades y oportunidades asociadas a **R**, y no repita errores parecidos, como el cometido a finales de los años ´90 e inicios de la pasada década, de distanciarse de un recurso metodológico tan potente y orientador, como es la *dialéctica-materialista*, para sucumbir a los *encantos* de los "vientos renovadores" de escuelas de la *metodología de la investigación científica* de dudoso sustento epistemológico, que lograron penetrarla en ocasión de la *apertura* y la atomización de aquel período.

Es imprescindible saber identificar lo útil y trascendente para el país, disponerse a aprehenderlo y masificar su empleo. En pocas cosas, como en la instrucción y en la formación de las nuevas generaciones, pueden los profesionales cubanos darse el lujo de improvisar o de andar a tientas, apartados del rigor y de la certeza que solo proporcionan la ciencia y la técnica verdaderas.

Referencias bibliográficas

Arencibia, R. & De Moya, F. (2008). La evaluación de la investigación científica: una aproximación teórica desde la Cienciometría. *Acimed 17(4)*, 1-27. Recuperado de http://scielo.sld.cu

Arnaiz, I. & García, J. A. (2011). La medición del impacto de la superación profesional de los docentes. Una alternativa para su perfeccionamiento. *Educación y Sociedad, 1-11*. Recuperado de http://www.revistaedusoc.rimed.cu

Bermúdez, R. & Rodríguez, M. (2008). Una aproximación más a la epistemología lógica y metodología de la investigación educacional. *Revista Pedagógica Universitaria, 13(3)*, 1-22. Recuperado de http://www.cvi.mes.edu.cu

Bliese, P. (2016). *Multilevel Modeling in R(2.6). A Brief Introduction to R, the multilevel package and the nlme package*.

Chirino, M. V. (2009). El método científico y los métodos de investigación en educación. En García-Batista, G. (Comp.), *El trabajo de diploma. Presentación oral y escrita* (pp.81-83). La Habana: Pueblo y Educación.

Cruz, M. (2009). *El método Delphi en las investigaciones educacionales*. La Habana: Academia.

Cruz, M., Escalona, M. & Téllez, L. (2014). Cualidad y cantidad en las investigaciones educacionales. Algunas reflexiones sobre su integración. *Revista Didascalia:*

Didáctica y Educación, 5(2), 203-222. Recuperado de http://revistas.ojs.es/index/php/didascalia/article

García, M., Montejo, M. N. & Carvajal, B. M. (2014). Estudio bibliométrico del contenido de ciencias pedagógicas en el corpus de la revista Transformación. *Revista Transformación, 10(2)*, 75-85. Recuperado de http://www.ucp.cm.rimed.cu/uzine/transformacion

Mainegra, D. & Miranda, J. (2012). Una propuesta para mejorar la comunicación de los resultados de la investigación educativa en la UCP ´Rafael María de Mendive´ en publicaciones de diverso formato. *Mendive, 41*. Recuperado de http://www.revistamendive.rimed.cu

Ortiz, E. (2015). Problemas que afectan la calidad de las tesis doctorales en Ciencias Pedagógicas. *Pedagogía Universitaria 20(2)*, 23-38.

Ortiz, E., González Guitián, M. V., Infante, I. & Viamontes, Y. (2010). Evaluación del impacto científico de las tesis doctorales en Ciencias Pedagógicas mediante indicadores cienciométricos. *Revista Española de Documentación Científica 2(33)*. Recuperado de http://redc.revistas.csic.es/index.php/redc/article/viewArticle/555

Pérez, L., Chávez, J, Rojas, C., Keeling, M., Piñón, J. C., Díaz, A., Rodríguez, M. A., Añorga, J. & Masó, R. (2009). *Aproximación al estudio de los aportes del ISP ´Enrique José Varona´ a la obra educacional de la Revolución Cubana*. La Habana: Academia.

R Core Team (2016). *An Introduction to R*. Disponible en http://127.0.0.1:11553/help/doc/manual/R-intro.html

Rubio, T. M. (2000). *Análisis de algunos indicadores bibliométricos aplicados a la revista Varona*. Recuperado de http://www.bibliociencias.cu/gsdl/collect/eventos/index/assoc/HASHO147.dir/doc.pdf

Torres, P. (2016). *Retos de la investigación educativa cubana actual. Aportes a su tratamiento*. La Habana, Cuba: Universidad en Ciencias Pedagógicas "Enrique J. varona". Recuperado de http://www.cubaeduca.cu/media/www.cubaeduca.cu/medias/evaluador/tesis2dogrado.pdf

Torres, P. & Lorenzo, R. (2015). ¿Cuán visible es la producción científica de la educación cubana actual? *Didascalia: Didáctica y Educación VI(4)*, 213-222.

Torricella, R. G., Van Hooydonk, G. & Araujo, J. A. (2000). Estudio bibliométrico sobre la presencia de los autores cubanos en el ´Web of Science´. *DataGramaZero. Revista de Ciência da Informaçâo, 1(4)*. Recuperado de http://www.brcpci.ufpr.br/documents.php?ddo=0000001218&dd1=5465a

Valledor, R. F., Ceballo, M., Blanco, M. & Ferrás, L. M. (2009). Una concepción de la investigación educacional. En García, G. Batista (Comp.), *El trabajo de diploma. Presentación oral y escrita* (pp. 94-106). La Habana: Pueblo y Educación

Recibido: 13 de enero de 2018
Evaluado: 25 de mayo de 2018
Aprobado para su publicación: 19 de julio de 2018